Thomas N. Friemel (Ed.)

Why Context Matters

Applications of Social Network Analysis

VS RESEARCH

Thomas N. Friemel (Ed.)

Why Context Matters

VS RESEARCH

Thomas N. Friemel (Ed.)

Why Context Matters

Applications of Social Network Analysis

VS RESEARCH

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

1st Edition 2008

Editorial Office: Christina M. Brian / Anita Wilke

VS Verlag für Sozialwissenschaften is part of the specialist publishing group
Springer Science+Business Media.
www.vs-verlag.de

Cover design: KünkelLopka Medienentwicklung, Heidelberg
Printed on acid-free paper
Printed in Germany

ISBN 978-3-531-16328-4

Content

Acknowledgements

This book compiles selected papers which were presented at the 4th conference on Applications of Social Network Analysis (ASNA 2007) at the University of Zurich. My thanks go to those associations and institutions which supported this event including the Swiss National Science Foundation, Hochschulstiftung der Universität Zürich, Vereinigung akademischer Mittelbau der Universität Zürich (VAUZ) and the host institute, the Institute of Mass Communication and Media Research at the University of Zurich (IPMZ). Special thanks go to the two keynote speakers of the conference, Tom A. B. Snijders and Noshir S. Contractor, as well as Christian E.G. Steglich, an inspiration for all participants' future research. Furthermore, I would like to thank the authors for their willingness to contribute to this book and the reviewers for their expertise and time invested. I must also express my gratitude to Heinz Bonfadelli who supported the conference and this publication in many respects.

More information about the conference can be found on the conference website: http://www.asna.ch or http://www.ipmz.uzh.ch/asna.

Thomas N. Friemel

Why Context Matters

Thomas N. Friemel

Applications of social network analysis are implicit answers to the question of whether context matters. The answer is consistently yes, context does matter. This finding is troublesome because it fundamentally questions the wide array of research which ignores context. But what is *context* and for what types of research questions does it matter? First, *context* is defined and its relevance to research is outlined. The second paragraph of this introduction gives an outlook on the various applications of social network analysis compiled in this book and groups them by application.

1 Re-incorporating context by applied social network analysis

Kurt Lewin stated that "every event depends upon the totality of the contemporary situation" (Lewin 1966: 10). Most of the time, however, the totality of a given situation cannot be captured in its entire complexity and reductions are unavoidable. Foremost, this holds true for quantitative research like that collected in this volume. Therefore, reductionism should not be questioned per se but is worth reflecting on. Reflecting critically is particularly important because reducing complexity is not always a fully systematic or intentional process. Often an incongruity between units of analysis and the units of recording (the units for which data is collected / observational units) exists. This gap is part of the definition of *context.* The *context* consists of all information which would be of interest to a research question less the information represented by the units of recording. Consequently, the question is not *if* context matters (because it matters by definition) but rather *why* context matters. It will be argued that the gap between the units of analysis and the units of recording is systematically biased and that social network analysis (SNA) is a powerful way to question and even overcome these biases and fill the gap by re-incorporating the context into the research setting.

In most scientific disciplines the units of recording are chosen by two guiding principles. On the one hand, the units are chosen according to naturally given enti-

ties. These are for instance human actors for which data is gathered in many social sciences, countries in political science and organisations in economics. On the other hand, the units are chosen according to the restrictions of data compilation. If primary data is collected, the unit needs to be able to react to an applied stimulus or expose the relevant information in a "readable" way. In social science, oral or written language is used as both a stimulus and reaction in the vast majority of research (i.e. people filling in a pencil and paper questionnaire). From a critical point of view, it can be questioned whether this technique is chosen because of its pertinence or rather because of its easy implementation. Especially when latent variables like feelings, personal traits, social groups, economic systems or political processes are the topic under study, appropriate alternatives are either not at hand or just difficult to realize.

Most often both mentioned factors – bias to (natural) given entities and data restrictions – privilege the same entities as units of recording. Hence, it is seldom questioned whether the chosen unit is meaningful and over time, well established standards have emerged in all research fields. For example, data is often gathered from individuals even though the research question addresses social groups and the true unit of analysis would therefore be an entity consisting of multiple individuals. By focusing on individuals as units of recording, the context (including the relations between the persons) is excluded from the analysis. Instead of incorporating context, most research systematically isolates the individual by applying random sampling. This is why Barton compared sample surveys to "a sociological meatgrinder, tearing the individual from his social context and guaranteeing that nobody in the study interacts with anyone else in it" (Barton 1968: 1). How absurd this is becomes apparent by the comparison with biology: "It is a little like a biologist putting his experimental animals through a hamburger machine and looking at every hundredth cell through a microscope; anatomy and physiology get lost, structure and function disappear, and one is left with cell biology" (Barton 1968: 1). This does not question the value of the micro-perspective in general but suggests that its scope is limited and that *context* does matter because the complex interactions across units of recording (cells or persons) in themselves constitute the unit of analysis (animals or social systems).

Social network analysis (SNA) is a way to re-incorporate context and bridge the gap between the micro and the macro, the cells constituting the animal, the individuals constituting groups, or the actors constituting a political system. SNA allows researchers to retain the traditional units of recording but simultaneously broadens the perspective by including information about the relationships across these units. This additional structural information allows researchers to address

existing research questions with new tools and to approach them from a different theoretical angle. The next section outlines eight selected examples.

2 Selected Applications of SNA

Applications of social network analysis can be divided in two groups of research: descriptive and explanative applications. *Descriptive applications* assess and describe the context by focusing on structural aspects. The question here is whether the observed structure is significantly different from a random structure. Measures for reciprocity, triadic structures, degree distribution and other regularity like multiplexity are at the core of this line of research. The contribution by *Victor Kryssanov et al.* is a prototype for this group. By analyzing the degree distribution, they take up an aspect of SNA which has been intensively investigated in recent years, foremost by physicists. A model framework is proposed to overcome the shortcomings of the existing power-law inspired approaches to modeling degree distributions of social networks. The proposed model is applied to datasets from patent authorship, paper citation, website visiting rate and delays in email reply. While this contribution focuses on the quantitative distribution of ties, the chapter by *Georg Müller* highlights the qualitative aspect of relations. He proposes a new methodology building on three-value logic of tie values. Beside the binary differentiation between true and false, a third value "possibly true" is introduced, which overcomes the inconsistent or contradictory coding that can emerge in binary logics. The technique is demonstrated in qualitative content analysis of semantic networks. The third contribution to the group of descriptive SNA is less methodologically and more substantially oriented. *Gerald Mollenhorst* addresses the sociological research question as to what extent people's public and private lives are two distinct spheres. This is tested by examining the congruency of social contexts in which people meet their acquaintances. In addition he analyses whether relationships are more uni- or multiplex. Analyzing a representative data set from the Netherlands, he finds differences for public and private contexts both for the structure overlap and the multiplexity of the relations.

The second group of studies can be labeled *explanative applications* because its focus lies on attributes which are used as dependent and independent variables in classic research. This line of research tries to explain how attributes of individual units are dependent on their structural embedding within a set of other units and their respective attributes. Likewise the causal order can be in the opposite direction where attributes are the independent variables which determine the structure

of the network. The first paper of this group addresses the prevailing topic of the semantic web. Based on the idea of two-mode networks, a rising amount of data is created on the internet linking people and objects. *Sheila Kinsella et al.* demonstrate how information from multiple online sources can be aggregated to a social network to highlight related people and objects. A second paper is also affiliated to the realm of communication and knowledge. *Wolfgang Sodeur and Volker Täube* address the question of how information flows in evolving social groups. In a first step, they look at social grouping among students and combine these findings in a second step with data on information exchange.

Centrality measures are among the most used concepts in SNA. This volume includes two contributions which apply centrality measures in connection with actor attributes. *Debra Hevenstone* analyses the relation between the prestige of academic departments' and their centrality in academic hiring networks. Her findings for US sociology departments suggest that this network resembles a positive feedback system. *Daniel Bochsler* shows that SNA can be combined with traditional statistical tools. He uses data from inter-governmental cooperation between Swiss cantons and uses the results from quadratic assignment procedure (QAP) and centrality measures in an OLS regression. Additional explanatory power of SNA is revealed applied in a longitudinal setting. Finally, *Shah Alam and Ruth Meyer* use agent-based simulations to describe the epidemic dynamics of HIV/AIDS within a population. This technique can help researchers design more precise interventions.

This compilation of applications of SNA includes theoretical, methodological and substantial advancements in a wide array of scientific fields and demonstrates the breadth of possible applications of SNA. It illustrates that SNA can approach long standing research questions more holistically, incorporating context into research. Classic research settings are extended and enriched by relational information. At the same time, SNA sheds light on new research areas (descriptive applications) which were completely ignored by previous research. With regard to the title of this book, we can conclude that for both groups of research *context matters*. Furthermore, it can be hypothesized that this research becomes ever more important as actors and objects which were once isolated become connected (Friemel 2007: 15). SNA, therefore, is not only an improvement to meet today's research challenge but also a promise for the future.

References

Barton, A. H. (1968). Bringing Society Back in. *American Behavioral Scientist, 12*(2), 1–9.
Friemel, T. N. (2007). Applications of Social Network Analysis (Introduction). In T. N. Friemel (Ed.). *Applications of Social Network Analysis*. (pp. 13–17). Konstanz: UVK.
Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1968). *The People's Choice* (3 [orig. 1944]). New York/London: Columbia University Press.
Lewin, K. (1966). *Principles of topological psychology [1936]*. New York.

A Hidden Variable Approach to Analyze "Hidden" Dynamics of Social Networks

Victor V. Kryssanov, Frank J. Rinaldo, Evgeny L. Kuleshov & Hitoshi Ogawa

Abstract This paper deals with the statistical analysis of social networks, and it consists of two parts. First, a survey of the existing, power-law -inspired approaches to the modeling of degree distributions of social networks is conducted. It is argued, with the support of a simple experiment, that these approaches can hardly accommodate and comprehensively explain the range of phenomena observed in empirical social networks. Second, an alternative modeling framework is presented. The observed, macro-level behavior of social networks is described in terms of the individual, "hidden" dynamics, and the necessary equations are given. It is demonstrated, via experiments, that a Laplace-Stieltjes hypertransform of the distribution function of human decision-making or reaction time often provides for an adequate model in statistical analysis of social systems. The study results are briefly discussed, and conclusions are drawn.

Acknowledgments The authors would like to acknowledge, with thanks, the use of the data collections received from Lada Adamic (Web-site hits) and Sune Lehmann (e-mail response statistics). We also thank Kohei Tsuda for his help in obtaining the patent authorship data. One of the authors (V.K.) acknowledges the financial support from Ritsumeikan University, the International Communication of Research Achievements Program.

1 Background and motivations

The abundance of the so-called "heavy-tailed" distributions, such as Pareto and Weibull ("stretched exponential"), in nature has created the present situation in empirical science, where the increasingly familiar scale-free pattern of the power (or Zipf, when dealing with rank statistics) law is eagerly discussed in disciplines spanning from physics and astronomy, through biology, linguistics, and economics, to sociology and computer science. This thus "ubiquitous" law is used to ana-

lyze and model such diverse phenomena as turbulent flow, size of naturally (e.g. physically, biologically, or socially) formed structures, word occurrence in a text, personal and corporate income, social (e.g. as of an author, actor, or a Web-site) popularity, private (e.g. communicative and sexual) contacts, and epidemic spread of infectious diseases (for more examples, see Newman, 2005, which also provides definitions of "scale-free" and other basic terms). In its simplest, "pure" form, the power law is expressed as $n_k = Ak^{-\gamma}$, where n_k is the number of discrete units of size k, $k = 1, 2, \ldots$, A is a normalizing term, the scaling exponent $\gamma > 0$, and it states that the corresponding probability mass (or density) function $f(k)$ of an observed random variable k follows (at least in the asymptotics) a hyperbolic form $f(k) \propto k^{-\gamma}$.

The outstanding attention to the power law in many currently "fashionable" fields investigating complex phenomena, such as development and evolution of social networks, may be attributed to three key factors: the simplicity of its formal definition, the transparency and convenience of manipulating the principal parameters of its possible generating mechanisms, and the apparent easiness of its detection in empirical data. Indeed, the power-law distribution has only one free parameter – γ, and its estimate $\hat{\gamma}$ obtained from data can often be used to completely characterize the underlying stochastic process, which presumably controls the random variable, and, hence, to predict the behavior of the observed system. Although there have been proposed a variety of stochastic processes generating this distribution, most of them can be classified into three groups (Mitzenmacher, 2003): power law through multiplicative bounded-minimum growth, power law through preferential attachment (or choice based on priorities), and as a result of optimization. These mechanisms were first explicated in the seminal works by Champernowne (1953), Simon (1955), and Mandelbrot (1960), in that order, and have since been thoroughly studied. Theoretically derived or calculated through simulation, values of the exponent γ may vary, depending on the underlying model chosen, but typically fall in the range from 1 to 3, while most of the relevant empirical data collections tend to reveal estimates of γ concentrated in a narrower interval somewhere in between 2 and 3 (Clauset et al., 2007). To detect the power law and obtain $\hat{\gamma}$ from data, a popular technique is to plot a histogram (often – with a magnitude-dependent width of bins) of the data on logarithmic scales that, in the power law case, should appear as a straight line. It is interesting to observe that while non-linear deviations in the tail areas of the histogram are very common in practice, power law seekers usually ignore them as long as some part of the binned data forms an "as-to-an-eye"-straight line.

Through a least-square (ordinary or weighted) regression analysis, an estimate of the slope of this line is then obtained that gives $\hat{\gamma}$. Amid the recently increasing criticism about the intrinsic inaccuracy of this "graphical" parameter estimation method, the maximum likelihood (MLE) technique is also used and provides for less biased and more robust results (Goldstein et al., 2004; Bauke, 2007).

Notwithstanding the over half-century -long study, the power-law behavior of various data is frequently presented as if not mysterious, quite controversial even in the specialized literature. There have been and still are vigorous debates about the "true" cause or originating mechanism of the phenomenon (e.g. see Adamic and Huberman, 2000, vs. Barabasi and Albert, 1999; or Stouffer et al., 2005, vs. Barabasi, 2005) and also about detection and parameter estimation technicalities (Clauset et al., 2007; Acosta et al., 2006). An overwhelming majority of recent high-profile publications on the subject deal with these issues. There exists yet a more fundamental problem: however surprising it may appear, there is no sufficiently (or at least comparatively) universal analytic alternative to the power law hypothesis proposed over the years. It is, apparently, this absence of alternatives that sometimes "forces" researchers to "decorate" their data (e.g. via preprocessing and consciously or otherwise selecting "illustrative examples") so that it would look as if revealing the scale-free behavior even when such behavior is subtle or not observed for most or some of the raw sample. The simplicity of the power law then turns up a huge disadvantage: having no other than γ as a meaningful parameter estimated from empirical data, which could independently (i.e. based on not merely assumptions made about the generally unobservable generating mechanism) be verified, makes validation of the model virtually impossible. To illustrate this criticism, let us conduct a simple experiment.

Fig. 1 displays results of the fitting of a sample comprising information about patent (co)authorship for $5\,517\,632$ inventors, whose names appeared in $15\,920\,951$ patent applications filed in various fields of manufacturing in Japan; the maximum number of (co)authored patents registered for one inventor is $1\,847$. The random variable k considered in the experiment reflects the number of patents registered for a particular inventor with the underlying social network having individual inventors as nodes and (co)authored patents – as (directed by the first author) ties (loops are allowed). Fig. 1 (A) suggests that from the two candidate-models (which are, in fact, the most popular degree-distribution models in social network analysis) of the corresponding "in-degree distribution," the power law may be selected as the "true" model, since its synthetic histogram (the solid line) well approximates the raw data behavior in a roughly two orders of magnitude range. Logarithmic binning of the data (Fig. 1, A, inset) would "improve" the fit

and further allow one to position the power law as a "reasonably accurate" model. From this point, a typical scenario would be to continue the analysis in the light of one or another well-studied power-law generating process, interpreting the calculated $\hat{\gamma} \approx 2.07$ (e.g. for the patent data, the "preferential attachment" idea would be brought in and appear very natural, out of plain "common-sense" considerations).

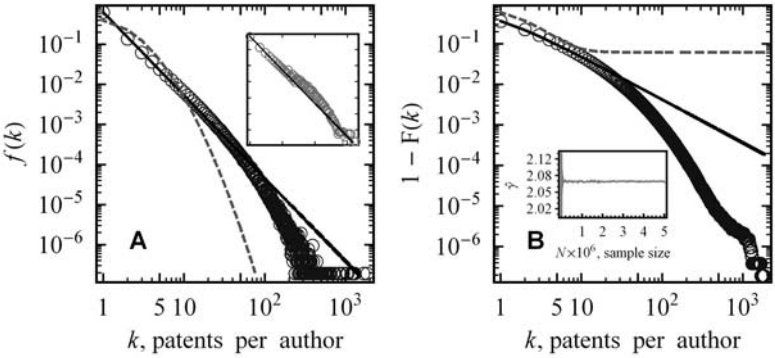


Figure 1: (A) The power-law (solid line) and lognormal (dashed line) models fitted (using MLE) to a patent authorship data sample (circles)

Inset (A): the same data logarithmically binned and the same power-law model. (B) The complementary cumulative plots for the same data and models; F(k) denotes the distribution function of k. Inset (B): values of $\hat{\gamma}$ calculated from different size datasets randomly sampled from the collection displayed in the main figure

Simple plots of the complementary cumulative sums of the real and synthetic data sets (Fig. 1, B) – a visualization technique surprisingly rarely employed in reports dealing with social network analysis – clearly indicate, however, that both the power-law and the lognormal models do not properly replicate the empirical distribution, especially in its right (heavy) tail. One would still argue that these deviations in the tail are caused by either an "underrepresentedness" of the empirical histogram for the large data values or the finite sample effects on model parameter estimation, and that increasing the sample size would result in improving the (tail) fit. In the considered case, however, the collection is much larger than normally used for a thorough statistical analysis in related domains, e.g. in sociology and economics. Furthermore, as shown in the inset of Fig. 1 (B), the calculated para-

meter value practically does not change (though not "by courtesy of" the model but is due to robustness of MLE) even for samples much smaller (actually, beginning from $N \sim 5000$) than the one used in the experiment. All this dictates that the power law is an inappropriate model for patent authorship (or inventors' productivity) – a conclusion, at which we could have arrived much earlier, should a relevant quantitative goodness-of-fit testing technique first be employed (e.g. Pearson's χ^2 test rejects both models of Fig. 1 at any sensible level of significance). The latter, unfortunately, remains also a rare practice in social network analysis (Clauset et al., 2007).

The lack of dissimilar analytic perspectives would typically restrict any subsequent data analysis to consideration of possible modifications of the power law "pure" form, e.g. by introducing additional model parameters, such as "shift," "percolation," or "cut-off" factors, while keeping the main focus on the "scaling" region(s) in the data behavior. (For instance, it would appear natural to break the raw data histogram shown in Fig. 1, A, into 2 parts with a cut-off point $k \sim 50$, so that each of the obtained segments would quite accurately be approximated with a straight line; for a practical example, see Lehmann et al., 2003). As it is often difficult to ascribe, based on domain or other "from-first principles" considerations, a physical (social, etc.) meaning to the thus ad hoc introduced parameters, computational experiments are usually conducted, aimed at reproducing the empirical data behavior and clarifying the role of the parameters in achieving this.

It should be noted, however, that in many cases of social network analysis (including the one of patent authorship), the scaling behavior is observed in areas best modeled with a value of γ less than 3 that assumes an infinite variance of the modeled sample mean. It may then be just meaningless to compare simulation results to the results calculated from the empirical data, because many parameters of the social network may, explicitly or implicitly, depend on the mean. Furthermore, model parameter values obtained via simulation may not be interpreted as a by-itself confirmation of a particular generating mechanism, unless the model is verified (at least, in part) with measurements other than those used to build it up. A proper model is also expected to be in agreement with facts about the modeled phenomena accumulated in related domains and, preferably, it should be founded on assumptions justified with arguments other than "common sense." The latter is a de facto standard in empirical science but is largely ignored in social network analysis (e.g. the notorious case of human response time – see Barabasi, 2005, but also Stouffer et al., 2005, and, as a classical monograph on the subject, Luce, 1986).

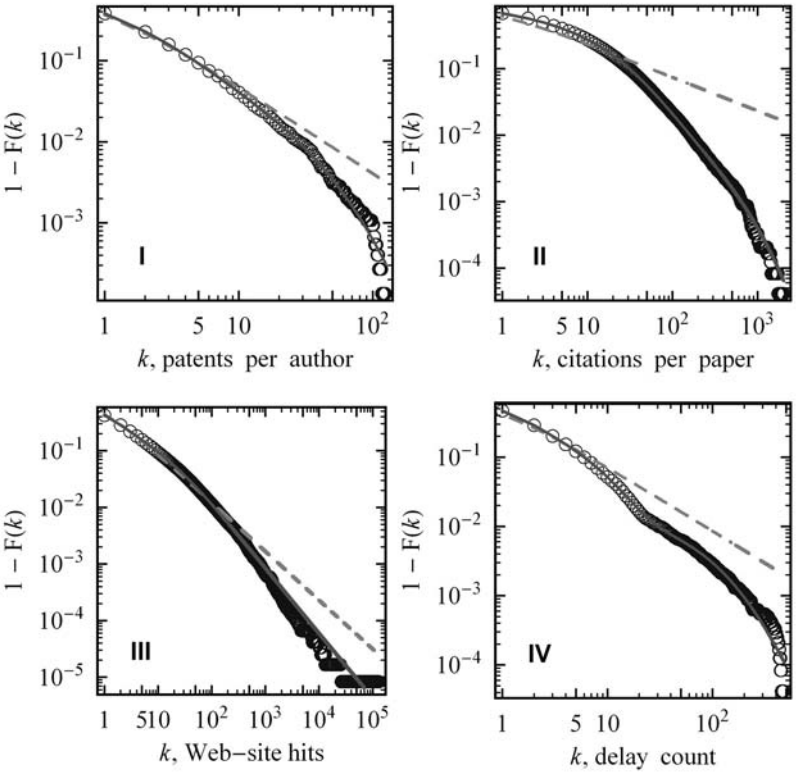


Figure 2: Results of fitting the functional forms of Eq. (2) derived in Section 2.4 to the data of (I) patent authorship, (II) paper citation, (III) Web-site popularity, and (IV) e-mail response time

In the graphs, circles represent experimental points, and solid light lines are theoretical curves with parameters calculated from the experimental data via numerical MLE. (For those curious, the dashed lines show the MLE fits of the (discrete) power law)

Summarizing the material of this section, we would like to point to three serious, in our opinion, problems in stochastic modeling of social networks. The well-studied power-law model may often offer a plausible explanation of the observed behavior in terms of some “basic,” “universal” mechanisms, but this simple model is (almost) never statistically sound. On the contrary, various modifications of the

power law may very well mimic the data behavior, but have obscure and questionable interpretations. In both cases, the models usually provide no facilities for independent verification and validation of their main theoretical assumptions. In this note, we would like to demonstrate that there does exist an alternative theoretical perspective for social network analysis, which addresses these problems and does not negate but instead generalizes the existing models. Fig. 2 is to illustrate right away the modeling accuracy achieved in the proposed framework while dealing with quite different data. In all the four presented collections, the data reflects certain aspects of the social macro-dynamics, which is naturally created out of the micro-dynamics of individual decisions and behavior. The following section then elaborates on the possible connections between these two dynamics and also seeks to establish validation arguments for the derived abstractions.

Before we proceed, it has to be noted that this paper is by no means intended to be a comprehensive theoretical discussion of the power law phenomena or a guide for their detection. Moreover, no graph-theoretic issues related to network dynamics and topologies are considered in this study, which mainly aims at statistical modeling and analysis of (various forms of) degree distributions. The interested reader can consult the relevant works (e.g. for the power law – Mitzenmacher, 2003; Newman, 2005; and Clauset et al., 2007, and for a graph-theoretic analytic perspective – Li, 2007).

2 Analytic framework

In this section, we will first formulate main assumptions of the proposed approach; a social network will be considered as a representation of the observed dynamics of a complex system. Next, we will argue that a standard exponential process is often a good model for describing the dynamics of one (fixed) state of a homogeneous (in respect to the generating mechanism and/or its mode of functioning) system. The exponential model will be extended to accommodate, first, the case of many-state dynamics and, second, inhomogeneous (whether temporally or physically) systems. This generalization yields a finite mixture of exponential-distribution-infinite-mixtures, which can be specialized to obtain the probability density function (PDF) of the random variable under analysis. Four examples of different social networks will then be considered, where parametric forms of the general model are derived by determining, based on domain considerations, the number of (homogeneous) subsystems contributing to the observed dynamics, as well as spectral (mixing over different states) distributions for each

of the subsystems. It will be demonstrated that the obtained models are not only more general, but also are much more precise and more convenient for analysis than models reported in the related literature. Some asymptotic properties of the derived distributions will be outlined.

2.1 Assumptions

We will consider a system Ω defined in a very general sense, i.e. as the object of investigation (not necessarily physically grounded). We will call an observable O a property of the system Ω that can be investigated in a given context. We will assume that Ω exists in different states, which release themselves as different outcomes of observations (e.g. measurements) associated with O. The latter means that the system states (or behavior, seen as state change) are in principle conceivable through their representations resulting from observations of O.

We will further assume that: *a*) there are multiple independent and competing observables for each state of the system; this means that at any time, only one but not necessary the same property of the system is evaluated – a conservation constraint for the observation (representation) time; *b)* the same state can have different representations; and *c*) different states can have the same representation.

In terms of social networks, nodes may often be thought of as systems, and ties – as representations of system states with the measured quantity defined as the node connectivity. In terms of statistical physics, a whole social network may be seen as a statistical ensemble of system states: a number of (imaginary) copies of one system (which are thus nodes) that are considered all at once, so that each copy is to represent a different possible (i.e. might-be-observed) state of the system.

2.2 One-state system dynamics

Let us denote ρ_i the representation change rate of the *i*-th observable of a system Ω, $i = 0,1,...,N$, with ρ_0 reserved for a given property O implicated in the measurement. Let us also denote ω the rate of one system state – ω is to characterize the internal, as opposed to the observed, dynamics of the system. Under the earlier made assumptions, the one-state observed dynamics of the system can be approximated by considering a “representational” diffusion process (see Kryssanov et al.,

2005, for the model differential equations) in the vicinity of a hyperplane $q\omega = \sum_i \rho_i$ formed by $N+1$ observables; q is a representation efficiency parameter indicating the extent to which the system state is in principle available for observation. Assuming that ρ_j, $j = 1,...,N$, are independent uniformly-distributed random variables, it is relatively straightforward to obtain that for a given observation time T, the distribution $\mathrm{F}(k)$ of $k = T\rho_0$ the count of different representations of the investigated state is well (for large N) approximated with a regular exponential process: $\mathrm{F}(k) = 1-(1-k/Tq\omega)^N \approx 1-e^{-k\beta}$, where $\beta > 0$ depends on the "hidden" (latent) parameters of the system and is mainly determined by the state average representation time. Alternatively, with fewer assumptions and somewhat more mathematically rigorous arguments, the exponential distribution for the one state representations can be derived within the Maximum Entropy statistical mechanics framework – see Kryssanov et al. (2006) for details on both methods. For the corresponding probability density and mass functions, we can now write:

$$f(k \mid \mathrm{B} = \beta) = \beta\, e^{-k\beta} \quad , \tag{1a}$$

and

$$f(k \mid \mathrm{B} = \beta) = (e^{\beta} - 1)\, e^{-k\beta} \quad , \tag{1b}$$

respectively, which thus specify the one-state dynamics, as reflected with the random variable k; k is continuous in Eq.(1a) and discrete – in Eq. (1b). Note that for the former, the observed state average rate $\langle k \rangle = 1/\beta$, and $\langle k \rangle \approx 1/\beta$ – for the latter.

2.3 The general case

Different states of the system under observation can naturally have different rates and, as a result, different representation times. For a statistical ensemble of states investigated by means of O and characterized by $\mathrm{F}(\beta)$ the cumulative distribution function (CDF) of the random variable β, the PDF (to keep technicalities low, we will mainly discuss the continuous case) of the state observed occurrences $f(k)$ is given by an infinite mixture of exponentials:

$f(k) = f(k \mid \mathrm{B} = \beta) \wedge_{\mathrm{B}} f(\beta) = \int_0^\infty \beta\, \mathrm{e}^{-\beta k} d\,\mathrm{F}(\beta)$. If the system is inhomogeneous, one would expect the existence of different $\mathrm{F}_i(\beta)$, $i = 1,2,...,M$, for its different subsystems and/or in different regimes of its functioning; $M-1$ is the number of inhomogeneities registered through O. This circumstance compels us to formulate the resultant model in a very general form as follows:

$$f(k) = \sum_{i=1}^{M} c_i \int_0^\infty \beta\, \mathrm{e}^{-\beta k} d\,\mathrm{F}_i(\beta), \tag{2}$$

where $c_1 + \cdots + c_M = 1$. Note, that Eq. (2) can be re-written as a Laplace-Stieltjes "hypertransform" $1 - \mathrm{F}(k) = \int_0^\infty \sum_{i=1}^{M} c_i\, \mathrm{e}^{-\beta k} d\,\mathrm{F}_i(\beta)$ with the left part easy to calculate from empirical data.

3 Application examples

Before we proceed with specific examples, there are several common issues worth to consider in the context of the proposed framework validation. Owing to Bernstein's theorem (Bernstein, 1928) and the fact that the PDF sought to replicate the observed system behavior is, in the case of social networks, usually a completely monotone function, Eq. (2) guarantees the existence of some proper $\mathrm{F}(\beta)$ called the spectral CDF (see Feldmann and Whitt, 1998), which describes the internal dynamics of the modeled (sub)system in terms of some "hidden" parameters. Although this latter dynamics is rarely available for observation, it is often possible to find out or at least conjecture about the distribution function of β, using knowledge of the domain under consideration. Domain knowledge may also be used for determining an optimal (parsimoniously) value of M, as there will always exist a risk of overfitting empirical data because of the excessive complexity of a particular parameterization of Eq. (2) chosen for tests. Besides, Akaike's Information Criterion, $\mathrm{AIC} = -2\log(\mathrm{L}(\hat{\theta} \mid x)) + 2n$, where $\log(\mathrm{L}(\hat{\theta} \mid x))$ is the log-likelihood maximized with a parameter set $\hat{\theta}$ of size n for a given sample x, can be used to estimate M (Akaike, 1983). By varying M, one should then select a model with the smallest value of AIC, as it will minimize the relative Kullback-Leibler distance between the model and the unknown true mechanism. In the following, we will mainly build on the relevant domain knowledge to achieve, at least in part, experimental control, to reduce chances of misinterpre-

tations, and to (again – in part) validate the analysis. The Akaike's Information Criterion has, however, also been used in all the experiments to ensure the choice of the simplest possible model.

3.1 Patent authorship: the Wald mixture of exponentials

Data used in this experiment is a sample representing the authorship of 9 349 patent applications filed by Japanese companies for 7 396 inventors in the field of micromachining during 1998-2003; the maximum number of patents (co)authored by one individual is 125. It has been learned through interviews with company managers that the applications may be associated with two disparate social "inventive mechanisms" quite standard for the industry in the inspected period of time (that is, however, not generally applicable to the much larger and different sample discussed in Section 1): R&D departments with long-term basic/strategic research projects and mass production with innovative activities. This knowledge can be used to justify setting $M = 2$ in Eq. (2) for the given sample of the patent authorship data.

As the next step in our analysis, it appears natural to assume that the productivity of an inventor depends on the rate of her or his problem-solving (decision-making). When working on similar (in a broad sense) problems, individuals with shorter, on average, problem-solving times are likely to succeed first and, hence, to get more patents. In other words, the count of patents per individual is inversely proportional to the time spent for solving the problem or, if put it in a more formal way, the human decision-making time and the parameter β of Eq. (2) are distributionally equivalent random variables. The sequential sampling evidence accrual mechanism (Luce and Raiffa, 1957; Luce, 1986) is one of the most-studied and well-grounded models of human decision-making used in experimental psychology and cognitive sciences. The latter model stipulates a Wald (Inverse Gaussian) distribution for the decision-making time (see Fewell, 2004) that allows us to utilize this distribution as a spectral CDF for our model (2): $\mathrm{F}(\beta) = (1 + \mathrm{Erf}[\sqrt{\lambda/(2\beta)}\,(\beta - \mu)/\mu] + \mathrm{e}^{2\lambda/\mu}\,\mathrm{Erfc}[\sqrt{\lambda/(2\beta)}\,(\beta + \mu)/\mu])/2$, where parameters $\lambda > 0, \mu > 0$, $\mathrm{Erf}[\cdot]$ is the (Gauss) error function, and $\mathrm{Erfc}[\cdot] = 1 - \mathrm{Erf}[\cdot]$ is the complementary error function. Through integration, Eq. (2) can now be specialized to a sum of two weighted (with c_1 and c_2) Wald mix-

tures of exponentials, each written as follows (component indices are omitted for clarity):

$$f(k) = \frac{e^{(\lambda - \sqrt{\lambda(2k\mu^2+\lambda)})/\mu}}{\sqrt{\frac{2k}{\lambda} + \frac{1}{\mu^2}}} \ . \tag{3}$$

Table 1: Discrete forms of the probability distributions used in the experiments (for parameter definitions, see the main text)

Distribution	Probability Mass Function	PDF (in the main text)
Wald mixture of exponentials	$e^{\lambda/\mu}(e^{-\sqrt{\lambda(2\mu^2 k+\lambda-2)}/\mu} - e^{-\sqrt{\lambda(2\mu^2 k+\lambda)}/\mu})$	Eq. (3)
Reciprocal-Wald mixture of exponentials	$e^{\lambda/\mu}\sqrt{\lambda}\left(\frac{e^{-\sqrt{\lambda(2k+\lambda-2)}/\mu}}{\sqrt{2k+\lambda-2}} - \frac{e^{-\sqrt{\lambda(2k+\lambda)}/\mu}}{\sqrt{2k+\lambda}}\right)$	Eq. (4)
Lomax (gamma mixture of exponentials)	$\big(b/(k+b-1)\big)^a - \big(b/(k+b)\big)^a$	Eq. (5)

Fig. 2 (I) defines k and shows results of the fitting of the form (2) with each summand set to the discrete analog of the distribution (3) (for the probability mass functions used in this and other three experiments, see Table 1), $M = 2$, and the parameter values calculated via numerical MLE as follows: $\hat{c}_1 = 0.83, \hat{\lambda}_1 = 1.43, \hat{\lambda}_2 = 0.17, \hat{\mu}_1 = 2.06$, and $\hat{\mu}_2 = 0.71$ (obviously, $\hat{c}_2 = 1 - \hat{c}_1 = 0.17$). Pearson's χ^2 test does not reject the model with a significance level $\alpha = 0.1$.

While a detailed analysis of the patent authorship process is not among the goals of this note, it is interesting to observe several facts, which immediately follow from the parameter (i.e. hidden variable) values obtained in the experiment. Specifically, one would contemplate that of the two social "patent-generating" mechanisms, the less (as estimated via $1/\langle\beta\rangle$, where the expectation $\langle\beta\rangle = \mu$ for the Wald-distributed random variable) "efficient" one (that is represented with the first component – weighted by c_1 and mainly for the law frequencies) is dominant

in terms of the people involved, and it can, with a high level of certainty, be attributed to the mass production ad hoc innovative activities. Asymptotic properties of Eq. (3) permit us also to speculate that in certain situations (specifically, when $\langle\beta\rangle^3/\langle\beta^2\rangle \to \infty$, $\langle\beta^2\rangle$ is the variance of β), it is unlikely to observe exceptionally prolific inventors, because patent generation is then a plain exponential process. Even from this rather superficial analysis, one may conclude that having a correct interpretation (e.g. in terms of the underlying social organization) for the hidden variables λ and μ could, perhaps, help optimize the inventive activities.

3.2 Paper citation: the reciprocal-Wald mixture of exponentials

In the second experiment, we explored citation statistics of 24 296 scientific articles published in Physical Review D in 1975-1994 and cited at least once; the maximum number of citations received by a single paper is 2 026, and the total number is 351 872. The data was obtained from http://physics.bu.edu/~redner/projects/citation/prd.html (last accessed on August 8, 2007). Beginning from this experiment, we will neglect interpretive aspects related to the structure of social networks underlying the data – the interested reader can find numerous examples of the relevant interpretations in the specialized literature, e.g. see works by Newman (2005) and Clauset with co-authors (2007).

In modeling the distribution of citations received by a paper, the key assertion will be that citation, similarly to invention and many other creative activities, depends on and is governed by human problem-solving. It appears natural to assume that the number of citations is directly proportional to the time expended (and, hence, to $1/\beta$) in the corresponding community to deal with (or "solve") the problem reported in the paper. This means that β of Eq. (2) is distributionally equivalent to a random variable with a reciprocal Wald distribution, if we again chose the Wald distribution for human decision-making time. The corresponding spectral CDF is defined as $\mathrm{F}(\beta) = \int_0^{\beta} (\sqrt{\lambda/\beta}\, e^{-\lambda(\beta\mu-1)^2/(2\beta\mu^2)} / \sqrt{2\pi})\, d\beta$ that, after some algebra, yields for each component of the resultant citation process model (2) the following specialization (the parameter indices are, again, not shown), where parameters $\lambda > 0$ and $\mu > 0$.

$$f(k) = \frac{\lambda\sqrt{2k+\lambda} + \mu\sqrt{\lambda}}{\mu(2k+\lambda)^{3/2}} e^{(\lambda - \sqrt{\lambda(2k+\lambda)})/\mu}, \quad (4)$$

Fig. 2 (II) depicts results of the modeling of k, the observed citation statistic with Eq. (2), where each summand is the discrete analog of Eq. (4), and $M = 2$ by assuming that citation of papers reporting on new problems is governed by social problem-solving with parameters significantly different from those in the case of reporting on well-recognized problems (e.g. owing to the so-called "social inertia" phenomenon, as discussed by Gerchak, 1984). These are the calculated parameter values (via numerical MLE): $\hat{c}_1 = 0.57$, $\hat{\lambda}_1 = 0.90$, $\hat{\lambda}_2 = 21.65$, $\hat{\mu}_1 = 12.22$, and $\hat{\mu}_2 = 16.37$. The model is not rejected by Pearson's test with a significance level $\alpha = 0.1$.

An important property of Eq. (4) is that its asymptotic forms include both an exponential distribution and a power law: $f(k) \to \mu^{-1} e^{-k/\mu}$ as $\lambda \to \infty$, but $f(k) \to \sqrt{\lambda}(2k+\lambda)^{-3/2}$ as $\mu \to \infty$. Various modifications of the latter models are habitually used in citation analysis (Bookstein, 1990; Lafouge, 2007). Another property is that for a paper, its chances to get ever cited grow (i.e. $f(0) \to 0$) only when both $\mu \to \infty$ and $\lambda \to \infty$. Having a plausible interpretation of the model parameters would, perhaps, help shed some light on why even high-quality and interesting reports may remain uncited, while other "not-so-good" articles are cited enthusiastically (see Kryssanov et al., 2007, that gives an example of a more detailed analysis of the citation process).

3.3 Web-surfing and reply-to-email delays: the gamma mixture of exponentials

In the next two experiments, we will deal with data – the Web-site visiting rate and the rate of responding to e-mails – representing human activities, which may all be characterized with one principal parameter called the human reaction time (RT). There are several alternative models of RT discussed in the specific literature, but the ex-Gaussian, lognormal, and gamma distributions are most commonly used to reproduce RT data obtained in experimental psychology and neurophysiology (Luce, 1986). Since in many situations, these distributions provide for roughly the same level of accuracy in predicting the human reaction time (van Zandt and Ratcliff, 1995; van Zandt, 2000), we will use the gamma distribution –

the form most convenient for the subsequent derivations. The spectral CDF for the data is, therefore, the gamma distribution $\mathrm{F}(\beta) = 1 - \Gamma(a, b\beta)/\Gamma(a)$, where parameters $a > 0, b > 0$, $\Gamma(\cdot)$ is the Euler gamma function, and $\Gamma(\cdot,\cdot)$ is the incomplete gamma function. The gamma mixture of exponentials is the well-known Lomax distribution (Harris, 1968) – a member of the Pareto family that can be written, in terms of Eq. (2), as follows:

$$f(k) = \frac{a\,b^a}{(k+b)^{a+1}} \,. \tag{5}$$

An important property of Eq. (5) is that for $k >> b$, it degenerates to the pure power law, and it can, thus, be considered as a generalization of this law.

Data used in the third experiment represents the Web-surfing activity of users of the America Online (AOL) Internet provider. The data sample was obtained from Lada Adamic (see Adamic and Huberman, 2000), and it covers statistics of accessing $119\,724$ sites by approximately $60\,000$ users during one day on December 1, 1997; the most visited site was accessed $129\,641$ times. Fig. 2 (III) defines k and displays results of the modeling of the site popularity with the model (2), where $M = 2$, summands are set to the discrete version of Eq. (5), and the parameters – to the following values obtained via numerical MLE: $\hat{c}_1 = 0.91, \hat{a}_1 = 1.07, \hat{a}_2 = 1.24, \hat{b}_1 = 0.67,$ and $\hat{b}_2 = 14.47$. The χ^2 test does not reject the model with a significance level $\alpha = 0.1$.

The main rationale for considering the reaction time as the variable controlling the Web-site visiting rate is that selecting (e.g. in a hypertext) or recalling a specific Web-site is a cognitive act universally characterized by a time expended to either find the relevant link(s) on the screen or to associate a theme or topic with a URL. It may be reasonable to expect that under other similar conditions, links with somewhat more "convenient," easy-to-find location/appearance on the screen would require, on average, less effort and shorter time to hit on, as few immediately perceived concepts can be stored in the working memory and are quick to retrieve and process (see Maljkovic and Martini, 2005). The latter, apparently, is also true for themes/topics and Web-cite addresses frequently experienced in one context in the recent past (Baddeley, 2007). Contrasting this, links not readily seen, as well as concepts with weak experiential connections require a longer time to process due to the unavoidable involvement of the long-term memory. In both cases, however, we have typical settings of (reaction time and recall) cognitive

experiments, in which empirical measurements are usually well emulated with the RT distributions (van Zandt and Ratcliff, 1995). Therefore, letting $M = 2$ and using the gamma distribution appears a justified choice for a candidate-model of the Web-site visiting rate. (Note that we do not hinge on the concept of Web-site popularity, as in the given context, "popularity" is not cause but consequence.)

Finally, Fig. 2 (IV) defines k and shows results of fitting the model (2) with summands specialized to the discrete analog of the distribution (5) to data representing the frequency of time-intervals between receiving an e-mail message and sending a reply (if any) to it (the discretization is 1 min). The individual delays were extracted from records of the timing of e-mails sent and received by approximately $10\,000$ people via a university network in Switzerland during a period of 81 days (which is also the longest delay); there are $23\,907$ e-mails replied. The e-mail traffic data was obtained from Jean-Pierre Eckmann (see Eckmann et al., 2004), and it was also used in other studies by Johansen (2004) and by Barabasi (2005). The original report pointed to the presence of a significant noise in the data, which would be attributed to technical factors (e.g. auto-reply by e-mail clients and processing long mailing lists by the servers). Furthermore, it was established through an analysis of the individual communications comprising the sample that there are two distinct, in terms of the response time, social groups exchanging e-mails (Eckmann et al., 2004). All this (a priory for the given study) knowledge influenced us to set $M = 3$ for the candidate-models. We, however, failed to find sufficiently convincing arguments for selecting a particular form of the spectral CDF and simply tried various combinations of the three distributions used in the previous experiments – the Wald, the reciprocal Wald, and the gamma functions. A rationalization (though admittedly weak) behind this is that any form of human communication can imply, if not be based on, recall and/or decision-making.

Among the models tested, the gamma spectral CDF for all the three components produced the best result (we have also tried this mixing distribution for a simpler model with $M = 2$ but did not obtain an acceptable fit). The model parameters were estimated (via numerical MLE) as follows:

$\hat{c}_1 = 0.01, \hat{c}_2 = 0.18, \hat{a}_1 = 4.24, \hat{a}_2 = 36.00, \hat{a}_3 = 2.73, \hat{b}_1 = 221.8, \hat{b}_2 = 206.2$ and $\hat{b}_3 = 2.34$. These figures are in a good agreement with our initial assumptions and results of Eckmann et al. (2004): the component indexed 1 can be associated with a high-frequency noise (in the range of "too-quick-to-be-human" responses). The subsystem modeled with the second (index 2) component is estimated as approximately 4 times smaller (or 4 times observationally less influential) but 6 times

more dynamic (and, perhaps, more constrained) than the subsystem modeled with the third (index 3) component; the subsystem sizes and dynamics were estimated via $\hat{c}_2$ and $\hat{c}_3 = 1 - (\hat{c}_1 + \hat{c}_2)$, and via the average response times defined in relative units as $\langle \beta \rangle = a / b$, respectively. The significance level α (with the χ^2 test) for this model is as low as 0.001 that may not be sufficient in certain situations.

4 Discussion: controlling the tail and other modeling opportunities

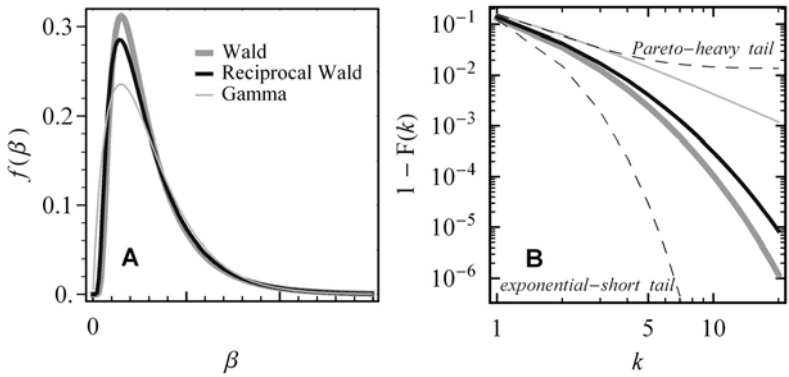


Figure 3: Different models of human micro-level dynamics (A) and the corresponding "degree-distributions" for the social macro-level (B)

The dashed lines display, for a reference, best MLE fits with the Pareto and the exponential distributions to the Gamma- and Wald- mixtures of exponentials, respectively. (The first and the second moments of the mixing distributions (A) are set to the same values)

The conducted experiments have demonstrated the flexibility of the proposed framework in modeling the observed behavior of different complex systems commonly identified as "social networks." Indeed, as it recently became evident (see section 1 and the literature cited therein), collective behavior statistics do not necessarily imply "pure" power-law or, otherwise, exponential tails and, hence, these statistics may not be replicated by the classic Pareto- or Poison- family distributions with accuracy sufficient for either analytic or practical purposes. This defi-

ciency is corrected in our model: Fig. 3 demonstrates how the right tail of the modeled degree distribution (pane B) depends on (is controlled by) the "hidden" dynamics of the observed system, i.e. the distribution of the average representation time of system states. Just as probability is shifted from around the mode to states with "unusually" longer (or shorter, when the time is inversely proportional to β) times in the mixing distributions (pane A), we move from the tail of the Wald mixture positioned in the middle of the exponential-to-Pareto "transitional" zone, to a semi-heavy tail for the reciprocal Wald, and to the nearly pure power law for the gamma mixture. Taking into account the fact that collective (macro) dynamics is formed (consciously or not) by individuals having their own (micro) dynamics, the presented three models appear quite natural and justified from the standpoint of cognitive psychology and behavioral science (see Luce, 1986, for the relevant models of what we call here the "micro-dynamics"). These are, though, not the only modeling opportunities.

Coming back to Eq. (2) derived in the beginning of this section, one may notice that for the specified exponential growth process, it will result in a log-normally distributed observed random variable, whenever the growth parameter is normally distributed (for a relevant example, see Vandermeer and Perfecto, 2006). This by no means original mechanism may easily be overlooked in a "traditional" power-law analysis but just pops up as, perhaps, the simplest idea to be tested first in the hidden-variable approach. (On the other hand, note also, that the gamma, the Wald, the log-normal, and other right-skewed distributions commonly used to model human behavior all converge to normality in asymptotics.) An analytic form, which is frequently discussed when empirical observations have "semi-heavy tails" and fall in a region somewhere close to but not quite in accordance with power law or log-normal model predictions, is the Weibull distribution function (Laherrere and Sornette, 1998). Equations (3) and (4) may be attributed to the Weibull family but, unlike the usual "stretched exponentials," they do not have specific cut-off points. If, however, the spectral CDF is defined, out of some domain considerations, as $\mathrm{F}(\beta) = 1 - \mathrm{Erf}[1/2\sqrt{b\beta}]$, which is a stable distribution, the PDF of the resultant mixture is $f(k) = \mathrm{e}^{-\sqrt{k/b}} / 2\sqrt{bk}$ that is a member of the Weibull distribution family with shape parameter $1/2$ (see Doyle et al., 1980, for related examples). Last but not least, it should be noted, that while it is generally (and strongly) advisable to knowledgably select spectral CDFs rather than try various functional forms merely because they were used in similar cases (e.g. as we did in the fourth experiment), any completely monotone heavy-tailed distribution can always be reproduced (with arbitrary precision) with a hyperexponential, i.e. a

finite mixture of exponentials (Feldmann and Whitt, 1998). This fact can be used to reconstruct from data, rather than to conjecture about, the spectral CDF describing the internal dynamics of a system – possibly, a new perspective in the context of social network analysis.

5 Concluding remarks

The analytic approach described in this paper has three strong points: I) stochastic models derived by specializing Eq. (2) are capable of thoroughly replicating degree distribution statistics of various social networks in most situations; II) such models can be verified, e.g. via the spectral CDF and its parameters, with results obtained in other domains, e.g. experimental psychology; and III) the popular distribution functions used in social network analysis are all assembled into one coherent framework as Laplace-Stieltjes (hyper)transforms of the functions representing the micro-level dynamics of the system under investigation. Discussing the approach weaknesses, one would point to the complexity of the models used in the experiments. Indeed, having to deal with 5-8 free parameters may not be very appealing in practice. It has to be remembered, however, that data of social networks (especially, those in large collections) can hardly be associated with a single, stationary and ergodic process: there may well be more than one generating process, and the process parameters may not remain the same throughout the data sample. Furthermore, even when the latter conditions are true, it may simply be incorrect to characterize the macro-behavior (i.e. of the network) with fewer parameters than required for models of the micro-level (i.e. human behavior), where 3-4 degrees of freedom is a norm (see van Zandt and Ratcliff, 1995).

The proposed framework cannot be called “entirely new:” there have been several reports discussing similar ideas for a statistical (e.g. Cohen, 1981; Abe and Thurner, 2005; Caldarelli et al., 2002; or more generally – Beck and Cohen, 2003) and even topological (e.g. Boguna and Pastor-Satorras, 2003) analysis of various systems. The main contribution of our work is that it combines many remote facts and seemingly unconnected studies into one systematic construction that, in fact, can be used as a practical guide for the modeling of a large class of complex phenomena.

For future work, and as a possible “remedy” for reducing the complexity of the very general form (2) obtained in this study, it appears interesting to consider more complicated and/or universal models of human decision-making and reac-

tion time, which would help us better understand the (un)observed inhomogeneities in the social dynamics.

References

Abe, S., Thurner, S. (2005). Analytic formula for hidden variable distribution: Complex networks arising from fluctuating random graphs. *Physical Review E*, 72, 036102.

Acosta, G., Grana, M., Pinasco, J.P. (2006). Condition numbers and scale free graphs. *The European Physical Journal B*, 53, 381-385.

Adamic, L.A., Huberman, B.A. (2000). Power-Law Distribution of the World Wide Web. *Science*, 287, 2115a.

Akaike, H. (1983). Information measures and model selection. *International Statistical Institute*, 44, 277-291.

Baddeley, A.D. (2007). *Working Memory, Thought, and Action*. New York: Oxford University Press.

Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, 435, 207-211.

Barabasi, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509.

Bauke, H. (2007). Parameter estimation for power-law tail distributions by maximum likelihood methods. arXiv:0704.1867v2 [cond-mat.other].

Beck, C., & Cohen, E.G.D. (2003). Superstatistics. *Physica A*, 322, 267-275.

Bernstein, S.N. (1928). Sur les functions absolument monotones. *ACTA Mathematica*, 51, 1 66.

Boguna, M., & Pastor-Satorras, R. (2003). Class of correlated random networks with hidden variables. *Physical Review E*, 68, 036112.

Bookstein, A. (1990). Informetric Distributions, Part I: Unified Overview. *Journal of the American Society for Information Science*, 41(5), 368-375.

Caldarelli, G., Capocci, A., De Los Rios, P., Mumoz, M.A. (2002). Scale-Free Networks from Varying Vertex Intrinsic Fitness. *Physical Review Letters*, 89, 258702.

Champernowne, D. (1953). A model of income distribution. *Economic Journal*, 63, 318-351.

Clauset, A., Shalizi, C.R., Newman, M.E.J. (2007). Power-law distributions in empirical data. arXiv:0706.1062v1 [physics.data-an].

Cohen, J.E. (1981). Publication Rate as a Function of Laboratory Size in Three Biomedical Research Institutions. *Scientometrics*, 3(6), 467-487.

Doyle, J., Hansen, E., McNolty, F. (1980). Properties of the mixed exponential failure process. *Technometrics*, 22, 555-565.

Eckmann, J.-P., Moses, E., Sergi, D. (2004). Entropy of dialogues creates coherent structures in e-mail traffic. *PNAS*, 101(40), 14333-14337.

Feldmann, A., & Whitt, W. (1998). Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31, 245-279.

Fewell, M.P. (2004). Comparative Descriptive Statistics of Skewed Probability Distributions, Technical report DSTO-TR-1596. Australian Government, Department of Defense (DSTO).

Gerchak, Y. (1984). Durations in Social States: Concepts of Inertia and Related Comparisons in Stochastic Models. *Sociological Methodology*, 14, 194-224.

Goldstein, M.L., Morris, S.A., Yen, G.G. (2004). Problems with fitting to the power-law distribution. *The European Physical Journal B*, 41(2), 255-258.

Harris, C.M. (1968). The Pareto Distribution As A Queue Service Discipline. *Operations Research*, 16, 307-313.

Johansen, A. (2004). Probing human response times. *Physica A*, 338, 286-291.

Kryssanov, V.V., Kakusho, K., Kuleshov, E.L., Minoh, M. (2005). Modeling hypermedia-based communication. *Information Sciences*, 174(1-2), 37-53.

Kryssanov, V.V., Kuleshov, E.L., Rinaldo, F.J., Ogawa, H. (2007). We cite as we communicate: A communication model for the citation process. arXiv:cs/0703115v2 [cs.DL].

Kryssanov, V.V., Rinaldo, F.J., Kuleshov, E.L., Ogawa, H. (2006). Modeling the Dynamics of Social Networks. arXiv:cs/0605101v1 [cs.CY].

Lafouge, T. (2007). The source-item coverage of the exponential function. *Journal of Informetrics*, 1(1), 59-67.

Laherrere, J., & Sornette, D. (1998). Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *The European Physical Journal B*, 2(4), 525-539.

Lehmann, S., Lautrup, B., Jackson, A.D. (2003). Citation networks in high energy physics. *Physical Review E*, 68(2), 026113.

Li, L. (2007). Topologies of Complex Networks: Functions and Structure. Ph.D. Dissertation. Pasadena, California: California Institute of Technology.

Luce, R.D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. New York: Oxford University Press.

Luce, R.D., & Raiffa, H. (1957). *Games and Decisions*. New York: Wiley.

Maljkovic, V., & Martini, P. (2005). Implicit short-term memory and event frequency effects in visual search. *Vision Research*, 45(21), 2831-2846.

Mandelbrot, B.B. (1960). On the theory of word frequencies and on related Markovian models of discourse. In R. Jakobson (Ed.), *Proceedings of the Twelfth Symposium in Applied Mathematics* (pp. 190-219). New York: American Mathematical Society.

Mitzenmacher, M. (2003). A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, 1, 226-251.

Newman, M.E.J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46, 323-351.

Simon, H.A. (1955). On a class of skew distribution functions. *Biometrika*, 42, 425-440.

Stouffer, D.B., Malmgren, R.D., Amaral, L.A.N. (2005). Comments on "The origin of bursts and heavy tails in human dynamics". arXiv:physics/0510216v1 [physics.data-an].

Vandermeer, J., & Perfecto, I. (2006). A Keystone Mutualism Drives Pattern in a Power Function. *Science*, 311, 1000-1002.

van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, 7, 424-465.

van Zandt, T., & Ratcliff, R. (1995). Statistical mimicking of reaction time data: Single-process models, parameter variability, and mixtures. *Psychonomic Bulletin & Review*, 2(1), 20-54.

Three-Valued Modal Logic for Reconstructing the Semantic Network Structure of a Corpus of Coded Texts

Georg P. Mueller

Abstract Researchers doing qualitative content analysis often face situations, where the preliminary results of the coding-process are inconsistent, too complex, and at a relatively low level of abstraction. This article presents a research method, by which these problems may successfully be tackled. It assumes, that the analyzed collection of texts is a manifestation of the cognitive map of the author of the texts, which in turn represents a relatively abstract, consistent, and parsimonious description of the author's environment. Hence, the article proposes to reconstruct this cognitive map by systematically transforming the initial coding of the texts. The article describes this process and presents appropriate heuristics for this purpose. At the end of the mentioned transformation of coded texts, there is a semantic network, which can be visualized by graphical means. Its nodes are the concepts of the described cognitive map and the links are implications from three-valued propositional logic. Hence, for describing cognitive maps, a third logical value „possibly true" is being used in this paper. This allows to define new types of implications between concepts, which are unknown in binary logic. Moreover, the third truth-value also helps to solve the mentioned problems with inconsistent or contradictory coding. The practical usefulness of three-valued logic for representing semantic networks of cognitive maps is illustrated by an example from a cooking book with Swiss specialties. The article analyzes the ingredients of typical Swiss soups and synthesizes the recipes in a semantic network structure, which is assumed to correspond to the cognitive map of the authors of the book.

1 Introduction

In the last few years there has been a growing interest in using computers not only for quantitative but also for qualitative content analyses of various kinds of texts and unstructured interviews (Fielding and Lee 1993, Kelle 1998, Kuckartz 2001, Miles and Huberman 2005, Lewins and Silver 2007). This trend has given rise to the development of new software products such as MAXqda, NVivo, NUD.IST, and ATLAS.ti, which can be used for automatic coding, text retrieval, hyperlinking of related text segments, etc. Some of these programs such as ATLAS.ti or MAXqda even allow to represent the results of qualitative content analyses in graphical form as semantic networks of coded texts (Sowa 1984: 76 ff., Lewins and Silver 2007: 179 ff.). Such networks consist of

1. text segments or so-called *quotations,* which generally constitute a non-overlapping partition of the analyzed text corpus,
2. *codes,* which are classificatory attributes of the mentioned text segments,
3. *links,* which are the result of the content analytic coding and describe the attribute relations between the mentioned codes and quotations.

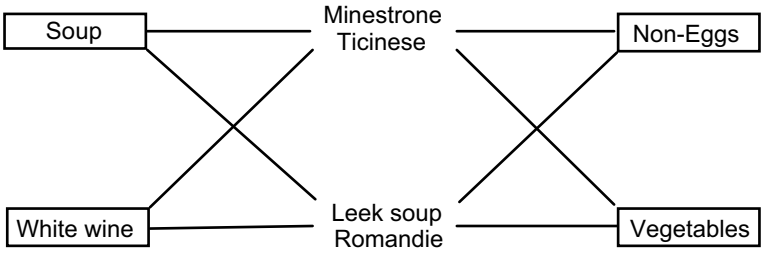


Figure 1: An example of a semantic network of a coded text: soup recipes from Latin Switzerland[1]

Fig. 1 gives a typical example of such a semantic network of coded texts, where the ingredients of soups from Latin Switzerland[2] are considered as codes or attributes of recipes, which represent text segments or quotations of the analyzed text corpus. As the example illustrates, semantic networks of coded texts have a

[1] For recipes and their ingredients see Klein and Tempelmann (2004: 43, 87).

[2] By „Latin Switzerland" we understand in this paper the French and Italian speaking regions of the country.

simpler structure than the *general* semantic networks, proposed by Sowa (1991) and others: there is only *one* type of attribute relation between codes and segments and the codes are never directly related. Moreover, fig. 1 also points to some potential *problems* of knowledge representation by semantic networks of coded texts:

1. For bigger text corpora, the number of coded text segments and attribute links rapidly grows and makes the *complexity* of the network less and less manageable.
2. Such network diagrams generally have a low *level of abstraction,* mainly due to the presence of coded text segments. This obviously makes theory construction rather difficult.
3. Due to contradictory text corpora or unreliable coders, coding is often *inconsistent:* in some cases quotations may simultaneously be coded by an attribute C and its negation ¯C = Non-C.

As a solution to the first two problems (1) and (2), the paper proposes to reconstruct from the coded text the so-called *cognitive map* of its author (Sowa 1991, Norman and Rumelhart 1978, Schank 1975: 164 ff.). This is the cognitive representation of that part of the external „world" of the author, which is described in the analyzed text. Thus, in this article, a cognitive map is a very general, abstract image of the environment of the author (Laszlo et al. 1996) and not only a cognitive representation of the spatial ordering of his/her surroundings (Kitchin and Freundschuh 2000). By reconstructing such a text generating cognitive map, we expect to get an abstract but not too complex summary of the essentials of the text. In the case of the recipes of fig. 1, the cognitive map, which we would like to reconstruct from the text corpus, corresponds to the general principles of cooking culture in Latin Switzerland.

In this paper, cognitive maps will be represented by systems of „if – then"-expressions from three-valued modal logic (Seiffert 1992, Snyers and Thayse 1989). Contrary to the classical binary logic, three-valued modal logic allows to reconcile the contradictory coding of text segments by attributing to some codes instead of „true" the third logical value, which expresses the *„possibility* of truth". This way, three-valued logic solves the previously mentioned problem (3) of inconsistent coding. Moreover, such systems of „if – then"-expressions have the advantage that they can be visualized as *semantic networks* with codes as nodes and links representing different types of implications from three-valued modal logic. Hence, in the following sections, this article describes the transformation process

leading from the original semantic network of a coded text to the semantic network of the text generating cognitive map.

2 The first steps from the coded text to the cognitive map

As already mentioned in the previous section, semantic networks of coded texts are complex systems of quotations, which are linked with one or several codes by attribute relations. As the high degree of complexity and the generally low level of abstraction are often due to the many quotations of such networks, it imposes itself to remove these text segments and to link the remaining codes *directly.* This is insofar a realistic approach as most networks of coded texts can be decomposed into typical triads of two codes A and B and a quotation in between (see fig. 2, above). From the perspective of propositional logic, A and B are propositions (= prop.) about the real world, which are simultaneously true such that from the upper part of fig. 2 follows A –> B and/or B –> A. As both inferences are the result of inductive reasoning based on the coding of just one quotation, a priori knowledge about causalities is useful in order to decide between these alternative inferences.

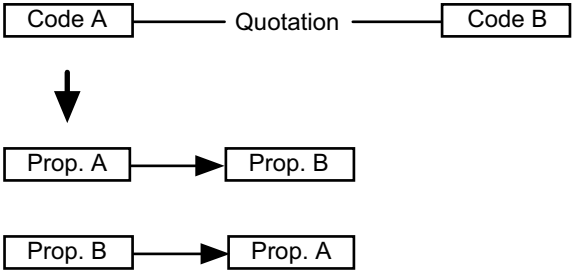


Figure 2: Logical inferences from triads of nodes in semantic networks of coded texts[3]

In sum, the first steps from a semantic network of a coded text to the semantic network of the text generating cognitive map are

1. to reinterpret codes as logical propositions,
2. to connect these propositions directly by appropriate logical implications,

[3] Prop. A and Prop. B are logical propositions corresponding to the codes A and B. The same meaning holds for all similar expressions in fig. 3 to 8.

3. to discard the quotations and attribute links, which are thus not needed anymore,
4. to discard redundant logical expressions, which often occur when treating many triads of codes and quotations.

Steps (2) and (3) are typical operations of qualitative co-occurrence studies in conventional content analysis (Krippendorff 2004: 205 ff., Neuendorf 2002: 175). They are often used in bibliometry (Callon et al. 1993: 78 ff.) in order to visualize by means of network-maps the proximity of scientific fields. The other steps (1) and (4) correspond to the simplest type of a configuration study in Qualitative Comparative Analysis (QCA) (Ragin 1998, Ragin 2000: chap. 3).

The afore-mentioned transformation process does generally not yield consistent, non-redundant, and complete cognitive maps. Sometimes the procedure rather reveals *contradictions* in the analyzed texts, which have to be reconciled by means from three-valued logic, that will be explained in the next following section 3. Moreover, there may still be a lot of unnecessary *complexity,* which can be reduced by generalizing secondary codes, as demonstrated later in section 4. Finally, the resulting semantic network may be *incomplete* with regard to its possible implications. This problem can partly be solved by adding complementary secondary codes. Partly it requires inference rules for three-valued logic, which will be a presented in section 5.

3 Three-valued logic for reconciling between contradictory propositions

3.1 An overview of old and new solutions

When transforming a semantic network of coded texts into a cognitive map, the first steps frequently lead to the contradictory situation described in the upper half of fig. 3: two quotations, which both belong to a code-category A, are coded as B and Non-B. In traditional binary logic the conclusion A –> B is in this case not possible, as in this logic A = "true" implies B = "true", which is obviously in contradiction with the fact that according to fig. 3 also ¯B = Non-B = "true".

There are several possible solutions to this inconsistency-problem. One of them is the identification of a mainstream coding and the definition of a *threshold* for acceptable shares of contradictions with regard to this coding. As long as the real shares of contradictions are below this level, the contradictory codes may simply be treated as errors of the coding process and consequently be dropped such that

the primary data become contradiction-free. A second solution is the use of *fuzzy logic* (Ragin 2000) for representing cognitive maps. As propositions in this type of logic have no absolute truth-value but rather a certain *probability* of being true, the coexistence of a proposition B and its negation ¯B is permissible.

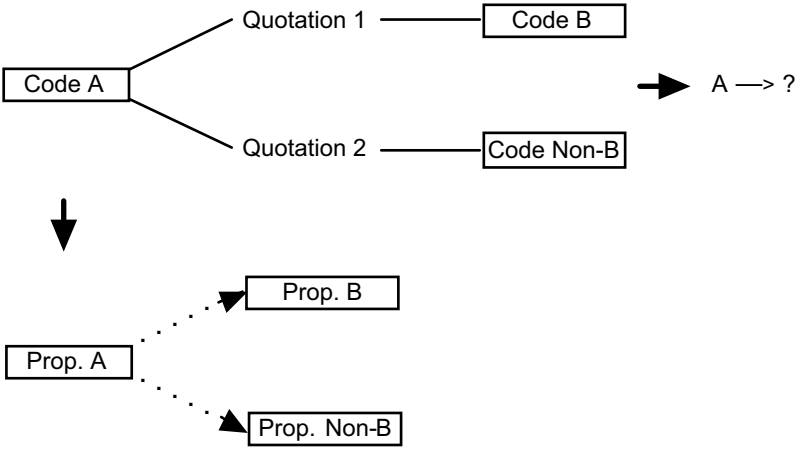


Figure 3: The use of *possible implications* as a solution to the problem of contradictory coding[4]

The aforementioned solutions are insofar not very satisfactory as they introduce *quantitative* elements such as thresholds or probabilities of truth into a type of text analysis, which is qualitatively oriented and which often aims at unconditional consistency, as exemplified by the goal of theoretical saturation in grounded theory (Strauss and Corbin 1998). Hence, in order to reconcile the aforementioned contradictions between B and ¯B, we propose to switch from traditional to three-valued modal logic (Rescher 1969: 22 ff., Malinowski 1993: 16 ff.), which was originally introduced by Lukasiewicz (1970 [1920]: 87, 88). Here we have as truth-values not only t = „true“ and f = „false“, but also p = “possibly true“ such that B and ¯B can *both* be consequences of the same true proposition A: if B is „possibly true“ we know from Lukasiewicz (see Rescher 1969: 23) that its negation ¯B is also „possibly true“. Hence, if A implies B with truth-value p = “possibly true“, then it becomes permissible that ¯B with the same truth-value p

[4] Prop. Non-B (=¬B) is the logical negation of proposition B. This definition also holds for all other subsequent figures. For the meaning of ¬B in three-valued logic see the appendix.

can also be a consequence of A, without running into logical contradictions (see dotted possible implications in fig. 3, below).

3.2 The possible implication in three-valued logic

By the introduction of a third truth-value p, it becomes possible to define new types of logical implications, which may be fruitful for a graphical description of the cognitive map of the author of a text. One of these new logical operators is the *possible implication* --->. Terms like A ---> B allow to express the idea, that there is the possibility that B follows from A. Hence, if A is true, B typically has a truth-value corresponding to at least p = „possibly true“ or even t = „true“. Other details of „--->“ are defined in tab. 1 by Lukasiewicz‘ implication =>, which is very similar to the implication –> in binary logic, and the possibility-operator ◊, which assigns ◊B the truth-value t = “true”, if B = p and thus only “possibly true” (see definitions of => and ◊ in the appendix). The possible implication ---> is typically used for reconciling *contradictory* or *inconsistent* coding like in the example of fig. 3. Besides, the possible implication ---> may also be used in situations of *uncertain* or *unreliable* coding, as exemplified by fig. 4.

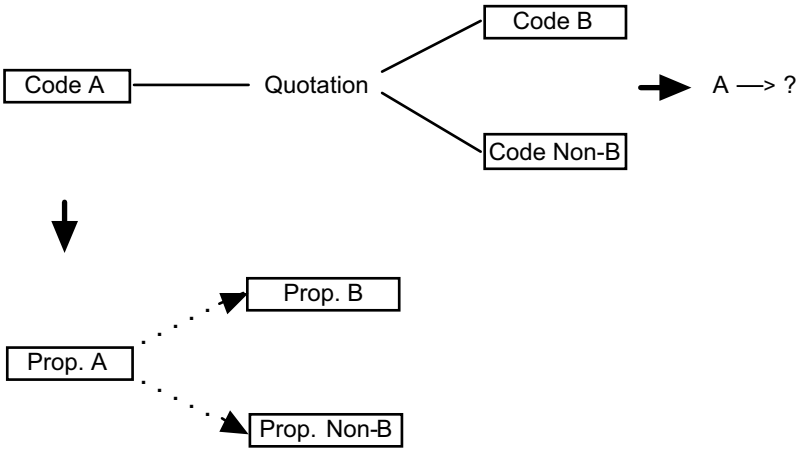


Figure 4: The use of *possible implications* for treating uncertain or unreliable coding

As in fig. 3 and 4 the implications A ---> B and A ---> ¬B are both true, the construction of the cognitive map may either be based on the more plausible or alternatively on both implications. In the end, the logical expressions of the resulting network-diagram must however be *true.* This has implications for the permissible valuations of A and B and the inferences that may be drawn from such a diagram (see section 5). E.g., according to tab. 1, only valuations of A and B represented by the lines 1, 2, 4, 5, 7, 8, and 9 yield an implication A ---> B which is logically true.

Table 1: The truth-table defining the possible, the necessary, and the inhibitory implication

No	A	B	A ---> B [A => ◊B]	A—> B [A => □B]	A -||-> B [A => □¬B]
1	t	t	**t**	**t**	f
2	t	p	**t**	f	f
3	t	f	f	f	**t**
4	p	t	**t**	**t**	p
5	p	p	**t**	p	p
6	p	f	p	p	**t**
7	f	t	**t**	**t**	**t**
8	f	p	**t**	**t**	**t**
9	f	f	**t**	**t**	**t**

--->: Possible implication. —>: Necessary implication. -||->: Inhibitory implication. [ ]: Equivalent expression in Lukasiewicz' logic (Rescher 1969: 22-25).[5] ◊: Lukasiewicz' possibility-operator. □: Lukasiewicz' necessity-operator. =>: Lukasiewicz' implication. ¬: Lukasiewicz' negation. p: Third truth-value, corresponding to Lukasiewicz' value „I". Bold **t**: "Flag" which marks the truth of an implication, as a requirement for representation in a semantic network diagram.

3.3 The necessary implication in three-valued logic

The second logical operator to be discussed in the context of three-valued logic is the *necessary implication* A —> B. It means that there is a necessity that B follows from A. This typically implies that from A = „true" follows that B must also be „true" and not only „possibly true", as in the case of the previous operator --->. For the other logical valuations of A and B, the truth of the expression A —> B is

[5] For definitions of the main operators of Lukasiewicz' three-valued logic, see truth-table in the appendix.

given in tab. 1. According to this table, the definition of „—>“ is based on Lukasiewicz‘ implication => and the necessity-operator □, which assigns □B the truth-value f = “false”, if B = p and thus only “possibly true” (see definitions of => and □ in the appendix). As the truth-table 1 shows after omitting all lines with A = p or B = p, the implication “—>” of three-valued logic is simply an extension of the implication “–>” of classical *binary* logic. Hence, „—>“ is mainly used in situations like fig. 5, where all relevant quotations have the same coding, such that there is absolutely no sign of any inconsistency and thus, in terms of binary logic, A –> B.

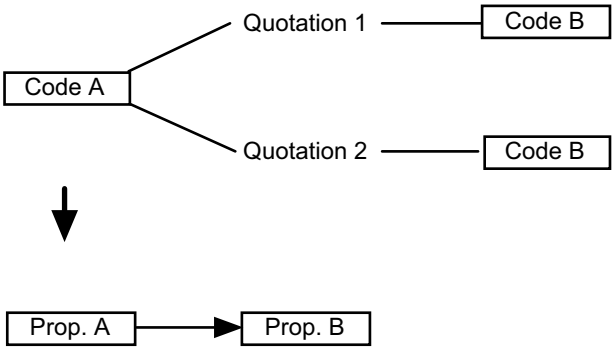


Figure 5: The use of *necessary implications* for treating quotations with identical positive coding

3.4 The inhibitory implication in three-valued logic

Finally, there is a third category of implications, which are also important for transforming coded texts into cognitive maps. In this article they are denoted by the symbol „-||->“ and called *inhibitory implications.* In expressions like A -||-> B, it is by definition inhibited that B follows from A. Hence, if A is „true“, the truth of the expression A -||-> B typically means that B must be „false“. Further details of the definition of A -||-> B are given in truth-table 1, where the operators ¬ and □ first invert the truth-values p and t of B into p and f, and subsequently transform both values into f (see definitions of => , ¬, and □ in the appendix). Inhibitory implications are typically used in situations like fig. 6, where quotations belonging to a code category A are consistently coded as ¬B: in this case, the universal prevalence of ¬B justifies the conclusion that there is an inhibitory implication A -||-> B.

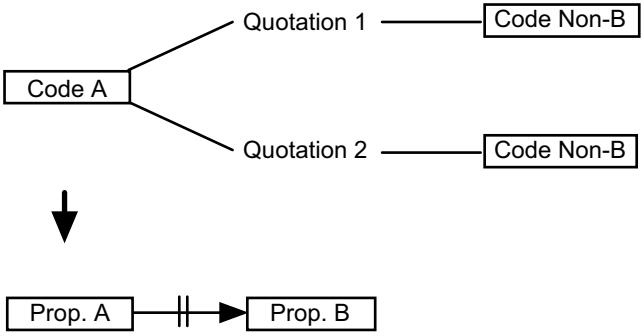


Figure 6: The use of *inhibitory implications* for treating quotations with identical negative coding

4 Secondary codes as general tools for reducing complexity and increasing completeness

The complexity of semantic networks of coded texts is not only due to the great number of quotations, but often also to the *scope* of the original coding scheme. As illustrated in the first part of fig. 7, this second type of complexity remains even after removing the quotations and linking their codes.

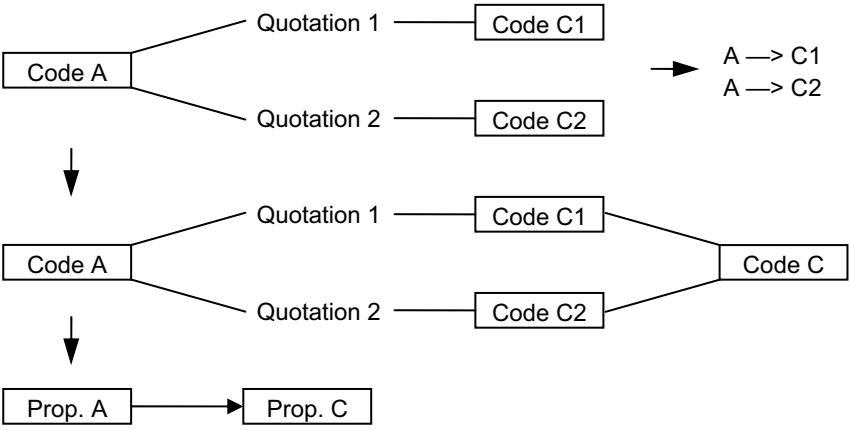


Figure 7: The use of a generalizing secondary code C

The solution proposed for this problem is rather general and consequently may be used for semantic networks based on three-valued logic: by the introduction of so-called *generalizing* secondary codes, such as e.g. generic terms, the original codes C1 and C2 can be bundled into an abstract meta-category C (see fig. 7). As the new secondary code C becomes this way a higher order property of the quotations 1 and 2, it can directly be linked to code A, such that finally A —> C (see fig. 7, below).

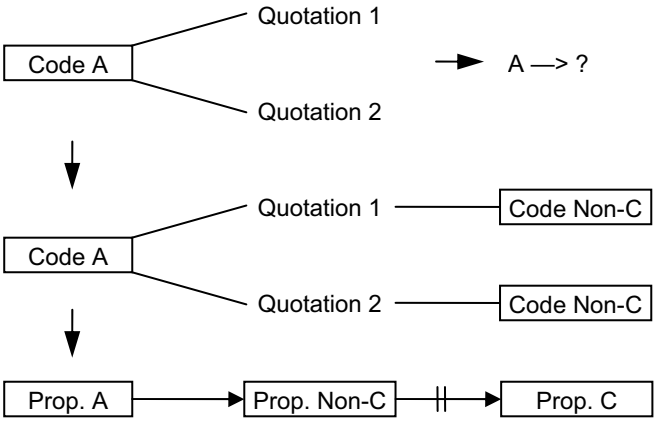


Figure 8: The use of complementary secondary codes ¬C = Non-C

Secondary codes are not only for reducing complexity: some of them may also be used in order to increase the *completeness* of a cognitive map, independently on whether it is represented in binary or three-valued logic. In particular, there are so-called *complementary* secondary codes, which are defined as logical negations of primary codes C. In certain cases, also the absence of a category C in the list of the codes attached to a given quotation may be interpreted as the presence of a complementary code ¬C = Non-C (see fig. 8). Such an interpretation makes e.g. sense in the case of the absence of certain ingredients in cooking recipes. As illustrated by fig. 8, complementary codes make our knowledge more complete and explicit, e.g. by indicating products, which must not be added to a recipe: in the latter case, a complementary code ¬C = Non-C and its complement ¬ ¬C = C are linked by an inhibitory implication "-||->" (see fig. 8).

5 Inference rules for increasing the completeness

One of the big advantages of logic as a language for describing the social world is the availability of precise inference rules, by which new propositions can be deduced from existing ones. This advantage obviously also holds for three-valued modal logic, which has its own, specific inference rules. By means of these rules, we may e.g. complete our knowledge about the cognitive map behind the analyzed text corpus. On the one hand, this additional knowledge is an end in itself. On the other hand, it is also useful for the empirical testing of models of cognitive maps. As they are always based on inductive generalizations from a particular text corpus, empirical tests may reveal that some of these models are not compatible with the available empirical facts.

One of the ways of creating additional knowledge about such semantic networks is the shortcutting of chains of logical implications by direct links between the start and the end of such chains. In two-valued propositional logic one would use for this purpose the transitivity-laws of logical implications. As we work here with three-valued modal logic with three different types of implications (see section 3), chains are more complicated and the mentioned transitivity-laws are not directly usable. Thus we are giving below some chaining-laws for three-valued logic, which can be proved by means of truth-table 1, either with paper and pencil or appropriate software:[6]

If „A —> B“ and „B —> C“ are both „true“, then „A —> C“ is also „true“.

If „A —> B“ and „B ---> C“ are both „true“, then „A ---> C“ is also „true“.

If „A ---> B“ and „B —> C“ are both „true“, then „A ---> C“ is also „true“.

If „A ---> B“ and „B ---> C“ are both „true“, then „A ---> C“ is also „true“.

In order to be able to use these chaining-laws also for expressions with inhibitory implications, we add two other inference rules, which too can be proved with truth-table 1:

[6] For complete proofs, up to 27 combinations of the three logical values of A, B, and C have to be analyzed by calculating the truth-values of the assumptions and implications of the investigated inference rules. It is possible to use for this purpose SPSS-data-tables with three truth-variables A, B, and C and 27 rows representing all logical value combinations of the mentioned three variables. By sorting of the SPSS-data-tables according to the values of A, B, and C, similar situations can be grouped together. This facilitates the calculation of the resulting truth-values of the assumptions and implications of the analyzed inference rules, which in turn can be represented as SPSS-variables.

If „A -||-> B“ is „true“, then „A —> ¯B“ is „true“, and vice-versa.

If „A —> B“ is „true“, then „A -||-> ¯B“ is „true“, and vice-versa.

Finally, for attenuating the strict implications —> and -||->, there are two other useful inference laws, which both follow from the definitions in tab. 1:

If „A —> B“ is „true“, then „A ---> B“ is „true“.

If „A -||-> B“ is „true“, then „A ---> ¯B“ is „true“.

Apart from the above mentioned inference rules there are obviously others. However, for completing the semantic networks of cognitive maps, the rules given in this section are almost always sufficient.

6 An exemplary application: Swiss soup recipes

In the following, we present an exemplary application of the methodology, which has been outlined in the previous sections of this paper. It refers to recipes of Swiss soups and asks about the typical ingredients of these dishes. The text corpus analyzed for this purpose consists of all soup recipes of the Swiss cooking-book of Klein and Tempelmann (2004), which we first have coded with regard to their ingredients. Fig. 9 gives a partial view of this coding.

Although Switzerland is a wine-producing country, wine is, according to fig. 9, *not* a universal ingredient of Swiss soups. Hence, in *two*-valued traditional logic, the proposition „Soup –> Red / White wine“ would be wrong. Three-valued logic, however, allows to hypothesize that red and white wine are both *possible* ingredients of Swiss soups (see fig. 10). By the virtue of the language of three-valued modal logic it is in fig. 10 also possible to postulate that the presence of white wine in a soup *inhibits* the use of red wine as an ingredient of the same soup, and vice-versa. Finally, fig. 9 also suggests that „Vegetables“ are codes attached to all Swiss soups such that in fig. 10 „Vegetables“ are a *necessary* implication of „Soup“.

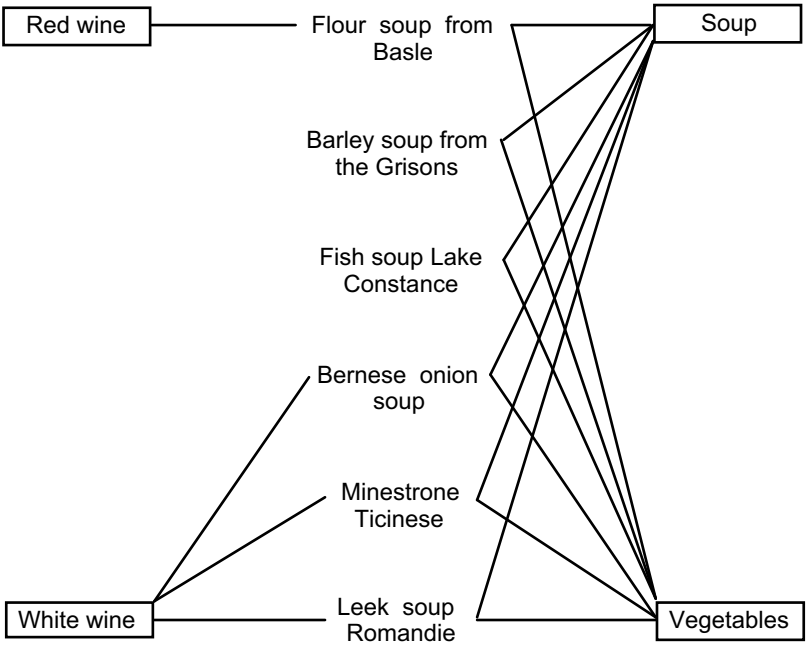


Figure 9: Swiss soup recipes as a semantic network of coded texts[7]

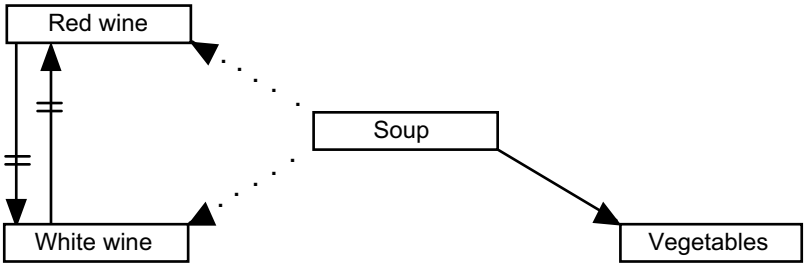


Figure 10: Possible, necessary, and inhibitory implications of soup recipes

In order to optimize the completeness and the complexity of fig. 10, we added to this cognitive map three secondary codes. (i) „Wine“ as a *generalizing* code,

[7] For recipes and their ingredients see Klein and Tempelmann (2004: 11, 23, 43, 53, 66, 87).

which is linked in fig. 11 by necessary implications to the subcategories „Red wine“ and „White wine“. (ii) „Non-Eggs“, which is a *complementary* code and stands for the absence of eggs in the list of ingredients of Swiss soups (see fig. 9). (iii) „Eggs“, which is the *complementary* code of „Non-Eggs“ and which is thus linked to the former category by an inhibitory implication.

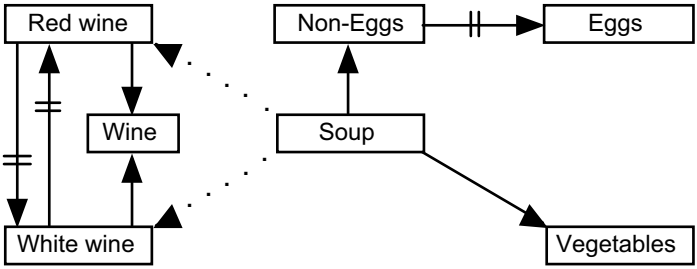


Figure 11: Soup recipes after the addition of all secondary codes

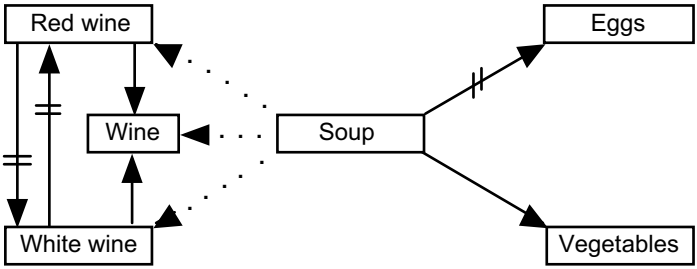


Figure 12: The completion of the cognitive map by logical inferencing

Adding new secondary codes to an existing semantic network generally opens the possibility of gaining new insights, just by applying the appropriate logical inference rules. Obviously, such conclusions can also be extracted from fig. 11 by using the logical laws of section 5. The results of this inference-step are given in fig. 12. It presents the cognitive map of the authors of the analyzed cooking-book as a semantic network, which is able to generate the original text coding of fig. 9. According to this updated map, a typical Swiss soup may possibly contain wine. It must however not contain any eggs (see fig. 12). This is a deductive conclusion from an *inductively* generated secondary code about the absence of eggs in the analyzed text corpus (see fig. 11). Consequently, it has to be empirically tested

with recipes from other cooking-books. If it turns out that in some of these new recipes eggs are an ingredient of Swiss soups, we would have to *revise* the cognitive map presented in fig. 12. As a matter of course, also other propositions of the final logical model have to be tested by additional new recipes, as they are too *inductive* generalizations from a particular text corpus.

7 Summary

Logical calculi have always claimed to give an adequate formalized description of the real world. This article stands in this tradition by using *three-valued modal logic* in order to reconstruct and describe the cognitive map of the authors of a coded text. This is insofar an innovation as three-valued logic is currently not very popular. For text- and content-analysis it is however useful, as it offers the possibility to handle contradictory and unreliable text-codes and to display their mutual relations in semantic networks with three different types of logical implications. The resulting maps are generally more detailed and more adequate than similar descriptions based on binary logic with only one implication. Fig. 13 clearly illustrates this advantage by an alternative cognitive map, which has been extracted from the coded text of fig. 9 by using only tools from *binary* logic. The comparison with the analogous map in fig. 12 demonstrates, how poor and incomplete this alternative map is – mainly due to the deficits of binary logic in grasping the essentials of the analyzed text.

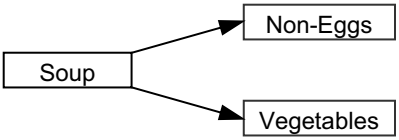


Figure 13: An alternative cognitive map based on binary logic[8]

[8] The propositions „Wine“, „Red wine“, and „White wine“ are not present in fig. 13, as the operators of binary logic do not allow to link them to „Soup“.

References

Callon, M., et al. (1993). *La scientometrie.* Paris: Presses Universitaires de France.

Fielding, N., & Lee, R. (Eds.) (1993). *Using Computers in Qualitative Research.* London: Sage Publications.

Kelle, U. (Ed.) (1998). *Computer-Aided Qualitative Data Analysis.* London: Sage Publications.

Kitchin, R., & Freundschuh, S. (Eds.) (2000). *Cognitive Mapping: Past, Present and Future.* London: Routledge.

Klein, M., & Tempelmann, Y. (2004). *Schweizer Küche, Cuisine Suisse, Swiss Cooking.* Lenzburg: Edition FONA.

Krippendorff, K. (2004). *Content Analysis.* Thousand Oaks: Sage Publications.

Kuckartz, M. (2001). *An Introduction to the Computer Analysis of Qualitative Data.* London: Sage Publications.

Laszlo, E., et al. (1996). *Changing Visions. Human Cognitive Maps: Past, Present, and Future.* London: Adamantine Press.

Lewins, A., & Silver, Ch. (2007). *Using Software in Qualitative Research.* London: Sage Publications.

Lukasiewicz, J. (1970) [1920]. *Selected Works* (ed. by L. Borkowski). Amsterdam: North-Holland.

Malinowski, G. (1993). *Many-Valued Logics.* Oxford: Clarendon Press.

Miles, M., & Huberman, M. (2005). *Qualitative Data Analysis.* Thousand Oaks: Sage Publications.

Neuendorf, K. (2002). *The Content Analysis Guidebook.* Thousand Oaks: Sage Publications.

Norman, D., & Rumelhart, D. (1978). *Explorations in Cognition.* San Francisco: Freeman and Company.

Ragin, Ch. (1998). Using Qualitative Comparative Analysis to Study Configurations. In U. Kelle (Ed.), *Computer-Aided Qualitative Data Analysis* (pp. 177-189). London: Sage Publications.

Ragin, Ch. (2000). *Fuzzy-Set Social Science.* Chicago: University of Chicago Press.

Rescher, N. (1969). *Many-valued Logic.* New York: McGraw-Hill.

Schank, R. (1975). *Conceptual Information Processing.* Amsterdam: North-Holland.

Seiffert, H. (1992). *Einführung in die Wissenschaftstheorie 3* (Kap. 2: „Modallogik"). München: Beck'sche Verlagsbuchhandlung.

Snyers, D., & Thayse, A. (1989). Modal Logic. In A. Thayse (Ed.), *From Modal Logic to Deductive Databases* (chap. 1.2). Chichester: John Wiley.

Sowa, J. (1984). *Conceptual Structures: Information Processing in Mind and Machine.* Reading (Mass.): Addison-Wesley.

Sowa, J. (Ed.) (1991). *Principles of Semantic Networks.* San Mateo: Morgan Kaufmann Publishers.

Strauss, A., & Corbin, J. (1998). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory.* Thousand Oaks: Sage Publications.

Appendix

Table 2: Definitions of the main operators of Lukasiewicz' three-valued logic

A	B	¬A	◊A	□A	A => B
t	t	f	t	t	t
t	p	-	-	-	p
t	f	-	-	-	f
p	t	p	t	f	t
p	p	-	-	-	t
p	f	-	-	-	p
f	t	t	f	f	t
f	p	-	-	-	t
f	f	-	-	-	t

(See Rescher 1969: 22-25)

Context Overlap and Multiplexity in Personal Relationships

Gerald Mollenhorst

Abstract In the sociological literature, it is suggested that grand changes in the structure of modern western societies in the 19th and 20th century, resulted in low levels of overlap among social contexts nowadays, which means that people usually meet each network member in a single social context. In this contribution, I examine the overlap structure among social contexts in which people meet personal network members, thereby especially focusing on overlap between public contexts and private contexts. Next, because it is also suggested that low levels of context overlap result in a replacement of multiplex relationships by uniplex relationships, I examine the extent to which sharing multiple contexts affects multiplexity in personal relationships. The main conclusions, which are based on empirical tests on data from the second wave of The Survey of the Social Networks of the Dutch, are a) that private contexts are much more likely to overlap with other contexts than public contexts, and b) that sharing multiple contexts in general, but especially sharing multiple private contexts, has a substantial positive effect on multiplexity.

Acknowledgments The author is grateful to Beate Völker, Henk Flap, Miller McPherson, several colleagues within the Interuniversity Center for Social Science Theory and Methodology (ICS), Thomas Friemel (editor of this volume), and two anonymous reviewers, for helpful suggestions and comments. This study is part of the program "Where friends are made. Contexts, Contacts, Consequences", which is supported by a Netherlands Organisation for Scientific Research (NWO) grant to Beate Völker as principal investigator.

1 Private and public: Two worlds apart?

It is suggested that grand changes in the economic and social structures of western countries in the nineteenth and twentieth century, like industrialization, the spread of wage labour, increased geographical mobility, and so on, resulted in a decrease in social cohesion and a loss of personal relationships. Communal solidarities of the past are replaced by a dominant concern for the private world nowadays and densely connected social networks degenerated into sparsely connected networks. And ultimately, people will no longer be members of a community, but 'bowling alone' (Putnam 2000). For examples of literature on this topic, see e.g., Stein (1960), Nisbet (1969), Fischer et al. (1977), Fischer (1982), Coleman (1990, 1993), Wellman (1999), and Pescosolido and Rubin (2000).[1]

From a sociological perspective, this alleged change in relational patterns is repeatedly explained by focusing on contextual opportunities and constraints. Coleman (1990, 1993), for example, speaks of an irreversible shift of activities that once took place in what he called primordial contexts, to activities nowadays taking place in purposive contexts; activities that formerly took place at home, at a relative's home or in the neighbourhood (like home care, child care, consumption of food, work and leisure activities) become more and more 'unbundled' and are taken over by modern, anonymous, social and economic organizations. Accordingly, people are expected to sacrifice multiplex relationships for single purpose relationships and social contexts no longer overlap. This lack of context overlap means that one works with one set of people, lives together with another set and spends his or her leisure time with a third set of people (Fischer et al. 1977, Coleman 1990, 1993).

This paper contributes to this field of research in two ways. First, I address the question to what extent people's public and private life have become two worlds

[1] Although these arguments have a long pedigree, there is little empirical evidence for this alleged decline in individual social capital. Wellman and Wortley (1990), for example, showed that most networks of East Yorkers (Toronto) have a 'saved' component as well as a 'liberated' component: "One segment of a network is composed of immediate kin whose relations are densely knit and broadly supportive, while other segments contain friends, neighbours, and workmates whose relations are sparsely knit, companionate, specialized in support, and connected with other social circles" (See also Wellman 1979). Hennig (2007), in a re-evaluation of the Community Question arrived at the same conclusion for German networks. And also recently, McPherson et al.(2006), concluded that "shifts in work, geographic, and recreational patterns may have combined to create a larger demarcation between a smaller core of very close confidant ties and a much larger array of less interconnected, more geographically dispersed, more unidimensional relationships".

apart, by examining the overlap structure among social contexts in which people nowadays meet their informal personal network members. Do people nowadays indeed meet their colleagues only at work? And do they meet their sports mates only at the sports club? Alternatively, do they meet the same people at both of these places? Or do they meet these people also at home? Second, I study whether meeting each other in multiple contexts is a condition for a relationship to be a multiplex relationship. Or the other way round, does meeting each network member in a single context mean that relationships are uniplex? These questions will subsequently be addressed in this paper.

2 Context overlap and multiplexity in informal personal relationships

When answering the aforementioned questions, I focus on informal personal networks, which consist of people who provide sociability, company, emotional support, and practical support. I study this part of people's personal networks, because it refers to that part of one's personal network that is not directly exogenously determined.[2] Moreover, it is less restricted than studies that focused on associates whom people discuss important matters with (e.g., Marsden 1987, 1990; McPherson et al. 2006). Examining personal relationships that provide practical help, emotional support, advice, or companionship, instead of just focusing on core discussion networks, therefore also provides a better picture of the current level of social isolation.

2.1 Context overlap

The overlap structure among social contexts can be examined by looking at the social contexts in which people currently meet their informal personal network members. I focus on meeting contexts like workplaces, neighbourhoods, voluntary associations and so forth, because these kinds of social contexts in which people meet each other are accurate measures of the 'foci of activity' through which personal relationships emerge and are maintained. According to Feld (1981:1016), "a focus is defined as a social, psychological, legal, or physical entity around which joint activities are organized […] As a consequence of interaction associated with

[2] See section 4.3.2

their joint activities, individuals whose activities are organized around the same focus will tend to become interpersonally tied and form a cluster." In others words, the social contexts (or foci) people enter in their daily life provide the pool of available others, out of which they can select personal network members.

Although social contexts can be categorized by a number of conceptual dimensions, like their size, degree of constraint, et cetera (see e.g., Coleman 1990: chapter 22, Feld 1981, and Fischer 1982), probably the most fundamental division of social contexts is that of public contexts, like workplaces, neighbourhoods, and clubs on the one hand, and private contexts, like one's own home, a relative's home or a friend's home on the other hand. It is this dimension that I use to hypothesize why overlap is more likely among some social contexts than among other contexts.

Since activities like home care, child care, food consumption, work, and leisure activities are more and more taken over by modern, anonymous, social and economic organizations (cf. Coleman 1993), people spend much more of their time in public social contexts than in private contexts. The fact that many of these public contexts are 'segmented' in one specific activity makes it unlikely that people meet the same group of people in all contexts. More specifically, the segmented character of public contexts makes that overlap among two or more public contexts is very unlikely. Somewhat more likely are overlaps between a public context and a private context. This means that people invite a colleague at home or meet their sports mate not only at a sports club, but also at home. Most likely, however, are overlaps among two private contexts. As soon as people meet each other at home, they are likely to meet each other at both their homes: relatives are likely to visit each other at home, and so do friends.

Hence the first hypothesis reads: *Context overlap is more likely between a private social context and a public social context as compared to two public social contexts, but most likely between two private contexts.*

2.2 Multiplexity and the effect of context overlap

Second, I examine the strength of the recurrently used argument for uniplex personal relationships. The aforementioned scholars commonly argue that the low degree of multiplexity in personal relationships is a result of people meeting each of their network members in a single social context. Although they seem to agree on this, this mechanism has not been empirically examined. This study therefore

addresses the question whether context overlap is a condition for relationships to be multiplex.

While the degree of multiplexity of relationships is addressed in various previous studies, it is also defined in various ways. In her research on multiplexity in adult friendships, Verbrugge (1979) already distinguished three definitions of multiplexity, as they were introduced by previous scholars. First, Gluckman, used the word multiplexity to refer to "the coexistence of different normative elements in a social relationship" (Gluckman 1962), which points at the coexistence of multiple roles within one relationship, for example, being colleagues as well as friends or relatives. So, a relationship is called uniplex if there exists just one role, and the more roles there exist, the more multiplex the relationship is called (see also, amongst others, Barnes 1969, 1972, Boissevain 1974, Fischer et al. 1977, and Marin 2004). Second, according to Kapferer (1969:213), multiplexity "simply refers to the number of exchange contents which exist in a relationship. In this case a relationship becomes multiplex when there is more than one exchange content within it, the minimum amount deemed necessary for a relationship to exist" (see also, amongst others, Mitchell 1969, Fischer 1982, and Haines and Hurlbert 1992).[3] And third, Wheeldon (1969:132) argued that "if the situations in which people habitually see one another are clearly distinguished it is possible to separate, very crudely, the strands which contribute to their relationship." His definition implies that a relationship is called 'multiplex' if people do not only see each other at work, but also belong to the same sports club. This definition is also used by Wellman and Wortley (1990).

Although I agree with Verbrugge (1979:1287) that "whether defined by roles, behaviours, or affiliations, multiplexity refers to *multiple bases for interactions* in a dyad", I think that there are still important differences between these definitions of multiplexity, especially between the first two on the one hand, as compared to the third definition on the other hand. Meeting each other in multiple social contexts (Wheeldon's definition), is one amongst other mechanisms through which the coexistence of multiple roles (Gluckman's definition), but in particular the coexistence of different activities or exchange contents in a relationship (Kapferer's definition) can occur. I therefore define multiplexity as the number of exchange contents in a personal relationship. This issue is previously addressed by Feld (1981:1024-25), where he stated, "a pair of individuals who share many foci are also likely to have multifaceted exchange relationships, but an analytical dis-

[3] A combination of both of these definitions is also used repeatedly, for example, by Ibarra (1995), Lazega and Pattison (1999), and Skvoretz and Agneessens (2007).

tinction should be maintained", by McPherson et al. (2001:437), and by Fischer et al. (1977:44-45).

Based on the structural argument that the grand changes in the structure of modern western societies resulted in non-overlapping (or unbundled) social contexts, which in turn resulted in a replacement of multiplex relationships by uniplex, single purpose relationships, the second hypothesis reads: *The smaller the number of social contexts in which informal network members meet each other, the less multiplex their relationship.*

3 Methods

3.1 The Survey of the Social Networks of the Dutch

In order to discern the overlap structure among social contexts for meeting informal personal relationships, and its effects on multiplexity in these relationships, I use data from the second wave of *The Survey of the Social Networks of the Dutch* (referred to as SSND2, see Völker et al. 2007). This second wave is a follow-up survey of the survey that was conducted in 1999/2000 (see Völker and Flap 2002). Seven years after the first wave, we re-interviewed as many of the original 1,007 respondents as possible. Over 70 percent of all respondents of whom we were able to retrieve their current home address were re-interviewed, which resulted in a dataset containing information on 604 individuals in The Netherlands, who are between 26 and 72 years of age.

Comparing these SSND2 data with national statistics on basic sociodemographic characteristics, we found that men, married people, older people and the higher educated were somewhat overrepresented. I therefore control for these personal characteristics in the analyses when possible.

3.2 Informal network delineation

The personal networks of the respondents were delineated through so-called 'name-generating' questions, 13 in total, which are presented in the appendix. Five name-generators asked for specific role-relations, like colleagues one frequently cooperates with and next-door neighbours (questions 3, 4, 5, 7, and 13). To deli-

neate informal personal networks, I do not make use of these role-related questions, because that part of one's network is directly exogenously determined. Instead, I focus on the remaining eight questions that generate people's 'informal personal network' (these questions are marked in the appendix with a star*).

Answering each of the name-generating questions, respondents were allowed to mention network members they had already mentioned in response to previous questions. In addition, they could add a certain maximum number of names every time (five in most cases, but the number is presented in the appendix directly after every question).[4]

Having collected the names of a respondent's personal contacts, additional questions (the 'name-interpreters') were posed on the relationship between the respondent and the network member. This provides the opportunity to control for the following variables in the analyses: frequency of contact, whether they like each other, duration of the relationship, and type of relationship.[5]

To determine the social context (or focus) in which people currently meet their informal network members, respondents were asked for every person mentioned: 'Where, on which occasion, do you meet person x nowadays?' They could choose up to three of the following contexts: 'at school', 'at a sports club', 'at a voluntary association', 'at another organization/ association', 'at work', 'at a relative's

[4] Only when answering question six, a sizable number of respondents named 5 new network members (about 10 percent), which indicates that restricting respondents to add 5 new network members per question maximally, hardly caused truncation of informal personal network size. Furthermore, it indicates that the number of network members that was named while answering the remaining five name-generating questions also hardly had any disturbing effect on the number of network members people could mention answering the eight 'informal network' name-generating questions.

[5] Frequency of contact was measured by asking 'How often do you usually have contact with person x?', with answer categories 'every day', 'every week', 'every month', 'every three months', 'once or a few times a year', and 'even less frequently'.

Liking each other is measured by asking 'Could you indicate, on a five-point-scale, to what extent you like person x?', with answer categories 'not', 'not much', 'somewhat', 'much', and 'very much'.

Duration of the relationship is measured by asking for the number of years they have already known each other.

Type of relationships is measured by asking 'How are you related to person x?', allowing respondents to choose up to three of the following categories: 'partner, living in', 'partner, not living in', 'parent, living in', 'parent, not living in', 'child, living in', 'child, not living in', 'in-law, living in', 'in-law, not living in', 'sibling, living in', 'sibling, not living in', 'other relative, living in', 'other relative, not living in', 'friend', 'boss', 'direct colleague', 'other colleague', 'former colleague', 'employee', 'someone from the neighborhood', 'direct neighbor', 'former neighbor', 'co-member of the same club/association', 'acquaintance', and 'other, namely……'.

home', 'at a friend's home', 'at my home', 'at their home', 'in the neighbourhood', 'at a public going-out place', 'at church', 'on a vacation, 'at a party', 'on the internet' and 'somewhere else'.

The variable on 'multiplexity' of a relationship measures the number of exchange contents in an informal personal relationship. This means that I counted the number of name-generating questions in reply to which each network member was mentioned by the respondent. To that aim, I combined the first and second name-generating questions into one exchange content 'giving and/or receiving advice', and the sixth and twelfth name-generating questions into one exchange content 'giving and/or receiving practical help'. This means that the multiplexity of a relationship ranges from 1 to 6.

3.3 Measuring overlap structure, using affiliation networks

The SSND data facilitate a reconstruction of the structure of social contexts for meeting network members and the extent to which they overlap. Usually, survey data order respondents in rows and their attributes (e.g., participation in events) in columns. An important property of these affiliation networks is that they allow us to study the dual perspectives of actors and events (Wasserman and Faust 1994). For the duality of relationships between actors and events, see Breiger's classic paper on 'the duality of persons and groups' (Breiger 1974).[6]

To examine overlap among social contexts for meeting informal network members, I started with a matrix, ordering network members (as they are mentioned by the respondents) in rows, and the social context in which they meet in columns. Next, I transformed this two-mode network into a one-mode network: I converted this 'network members by contexts' matrix to a 'contexts by contexts' matrix, by multiplying the transpose of the 'network members by contexts' matrix with the 'network members by contexts' matrix itself[7], using the software package Ucinet (Borgatti et al. 2002). The cells of the resulting sociomatrix (Table 3) present the degree of overlap among social contexts for meeting informal personal network members. To illustrate, consider a person who meets a network member at school as well as at work. This contributes 1 to the overlap frequency for 'at

[6] Examples of research that have employed affiliation networks are enumerated by Wassermann and Faust (1994:295-6).

[7] For a mathematical explanation of this approach, see for example, Wasserman and Faust 1994: chapter 8.

school' versus 'at work' in the matrix. Consider a second relationship with a person she/he meets at a sports club, at work, and at church. This contributes 1 to the overlap frequency for 'at a sports club' versus 'at work', 1 to the overlap frequency for 'at a sports club' versus 'at church', and 1 to the overlap frequency for 'at work' versus 'at church'. The numbers on the main diagonal in this matrix show the total number of 'memberships' in each of the social contexts, i.e., the total sum of network members of all respondents altogether who meet in each of the social contexts.

Although the frequencies of overlaps among meeting contexts of network members are informative in themselves, their description of overlap is distorted by the 'size' of each context. Two popular contexts for meeting network members show a higher overlap than two non-popular contexts, not necessarily because these contexts 'appeal' to each other, but simply because in both contexts many people meet each other (see, e.g., Bonacich 1972, Faust and Romney 1985, Wasserman and Faust 1994). To control for this 'context size', I use Bonacich' (1972) normalization method. The resulting normalized matrix (not presented) is then used as input to a Johnson's (1967) hierarchical clustering analysis. Both methods are implemented in Ucinet (Borgatti et al. 2002). The results of this analysis are presented in Figure 1.

3.4 Multivariate multilevel analysis

The final part of my examination consists of a multivariate multilevel analysis on the effect of context overlap on multiplexity in informal personal relationships (Table 5). The dependent variable 'multiplexity' measures the number of exchange contents in an informal personal relationship. In all four models in this table, I control for respondent's age, sex, level of education, marital status, employment status, and network size.

In the first model, I examine the effect of context overlap, using two dummy-coded variables which indicate whether the respondent and the network member meet each other in two or in three different social contexts. The number of contexts is split up further in the second model, indicating whether the overlap involves 'private contexts' or 'public contexts'. In this model, the contexts 'at a relative's home', 'at a friend's home', 'at my home', and 'at their home' are considered 'private contexts', whereas all other contexts are considered 'public contexts'.

In the third model, I add four relationship characteristics, which are expected to have an effect on multiplexity: frequency of contact, the extent to which they like each other, the duration of the relationship, and the type of relationship (see, e.g., Fischer et al. 1977:44-45). To that, the original variable on contact frequency is recoded into two dummy-coded variables for 'weekly contact', and 'daily contact', such that relationships with contact frequencies less than once a week make up the reference category. The initial variables on types of relationships were recoded into seven dummy-coded variables for 'partner', 'relative, living in', 'relative, not living in', 'friend', 'co-worker', 'neighbour', and 'co-member', such that those who do not fit into one of these categories (e.g., acquaintances) make up the reference category.

4 Results

In this section, I present results of empirical tests of the hypotheses on the overlap structure among social contexts and on the effect of context overlap on multiplexity in informal personal relationships (sections 4.2, respectively section 4.3). Before that, I describe the size and composition of informal personal networks and the multiplexity of these relationships in section 4.1.

4.1 Size and composition of informal personal networks

Table 1 shows that the average informal personal network consists of about ten members, although there is substantial variation among respondents. About 25 percent of the respondents reported having six network members or less, while almost seven percent mentioned merely three members or less. In substantive terms, this means that one of four people has no more than six different people with whom she or he discusses work-related or personal problems, whom she/he pays a visit sometimes, from whom she/he receives practical help with odd jobs in or around the house and with whom she/he sometimes spends a night out. A considerable number, about 40 percent of the respondents, however, reported having over ten network members with whom they are involved in these types of activities.

Table 1: Informal personal network size and compositiona

Variable	#	%	Mean	St.Dev.	Mode	(N)
Overall informal personal network size			9.75	4.65	9.0	604
0 – 3	41	6.8				
4 – 6	115	19.0				
7 – 10	210	34.8				
11 – 15	170	28.2				
16 or more	68	11.3				
	604	100.0				
Types of relationships						604
partner			0.67	0.57		
relatives			2.68	2.23		
friends			2.28	2.10		
co-workers			2.01	2.27		
neighbours			1.61	1.57		
others[b]			1.48	2.32		
Number of types of relationships			*1.10*	*0.33*		*5,894*

Source: SSND2, 2007
a) Eight name-generators were used to delineate these networks (see Appendix);
b) 'No answer' is included in this category.

On average, one third of all informal personal network members are relatives, partners included ([2.68+0.67]/9.75). About a quarter are called friends (2.28/9.75), a surprisingly small part, given the content of the name generating questions we used. It actually means that over 40 percent of one's informal network members are not one's partner, relative, or friend, but are 'just' called a co-worker, neighbour, co-member of the same club, acquaintance or whatsoever ([2.01+1.61+1.48]/9.75). Moreover, respondents predominantly have single role relationships with their informal network members (1.10 on average).

The first part of Table 2 shows that informal network members are most relevant when it comes to giving and/or receiving a hand with odd jobs at home or to paying a visit; people on average report having more than four associates for each of these activities (4.33, respectively 4.36). Moreover, they also mention having three people who will help them in case they fall ill. The average size of the 'core discussion network' is 2.41 (i.e., the number of network members with whom one discusses important personal matters). This number stayed remarkably stable since the previous wave of our survey in 1999/2000 in which we found an average size

Table 2: Exchange contents and meeting contexts

Variable	Mean	St.Dev.	(N)
Exchange contents in relationships			604
give or receive work-related advice	1.90	2.22	
give or receive help with odd jobs	4.33	2.80	
pay him/her a visit	4.36	2.64	
discuss important personal matters with	2.41	2.21	
going-out with	2.51	2.37	
provide me with help when I'm ill	3.04	2.40	
Multiplexity[a]	*1.90*	*1.11*	*5,894*
Contexts in which they currently meet			604
at school	0.04	0.21	
at a sports club	0.33	0.73	
at a voluntary association	0.14	0.55	
at another association	0.25	0.77	
at work	2.01	2.30	
with family	1.60	1.83	
with friends	0.81	1.37	
at respondent's home	5.97	3.54	
at the network member's home	5.02	3.15	
in the neighbourhood	1.24	1.41	
at a going-out place	0.38	1.11	
at church	0.15	0.61	
on a vacation	0.31	0.72	
at a party	0.75	1.50	
on the internet	0.07	0.39	
somewhere else[b]	0.61	1.88	
Context overlap[c]	*2.02*	*0.89*	*5,894*

Source: SSND2, 2007

a) Multiplexity is the number of exchange contents within a relationship;

b) 'No answer' is included in this category;

c) Context overlap is the number of contexts in which they currently meet each other.

of 2.37 (Mollenhorst et al. 2008a). Multiplex relationships are not rare: The average number of exchange contents in a relationship is 1.90. Additional analyses, which are not presented in these tables, showed a) that people are likely to discuss

their important personal matters with people they also go out with, b) that those who help each other with odd jobs around the house are also likely to give a hand when one falls ill, and c) that relationships with people who are asked for or who give advice on problems at work are hardly ever multiplex relationships.The second part of Table 2 shows where informal network members currently meet each other. The vast majority meet at home: An average number of six informal members meet at respondent's home, whereas five meet at the network member's home. Of course, household members of the respondent mainly cause the difference between these two numbers. The next most important social contexts to meet informal network members are the workplace (where on average 2.0 members are met), a relative's home (1.6 on average), and the neighbourhood (1.2 on average). All other social contexts are of minor importance for meeting informal network members. Finally, this table shows that people on average share two contexts with their network members.

4.2 Overlap among meeting contexts

In Table 2, I showed where informal network members currently meet each other, and that they share two social contexts on average. Next, Table 3 presents the overlaps among these social contexts. To give an example: The figure in the first row and first column shows that 3,605 (of all 5,894) network members are met at respondent's home. Next, the figure in the first row, second column (which is the same as the figure in the second row, first column) shows that 2,790 network members are met both at respondent's home as well as at the network member's home. The off-diagonal figures therefore indicate the overlap among social contexts.

Table 3 shows that the by far most frequently occurring context overlap is between respondent's home and the network member's home (2,790 instances), which actually makes up 29.6 percent (2,790/8,435) of all overlaps. The second most frequently occurring overlap is between respondent's home and a relative's home (794 instances). Looking more broadly, one can see that most overlaps occur among the 'private' contexts respondent's home, the network member's home, a relative's home, and a friend's home. Altogether, overlaps among these contexts make up 59.2 percent of all overlaps (4,993/8,435). Next, overlaps in which one private context and one public context are involved make up 35.5 percent of the overlaps (2,993/8,435), which means that just 5.3 percent of all overlaps occur between two public contexts (449/8,435).

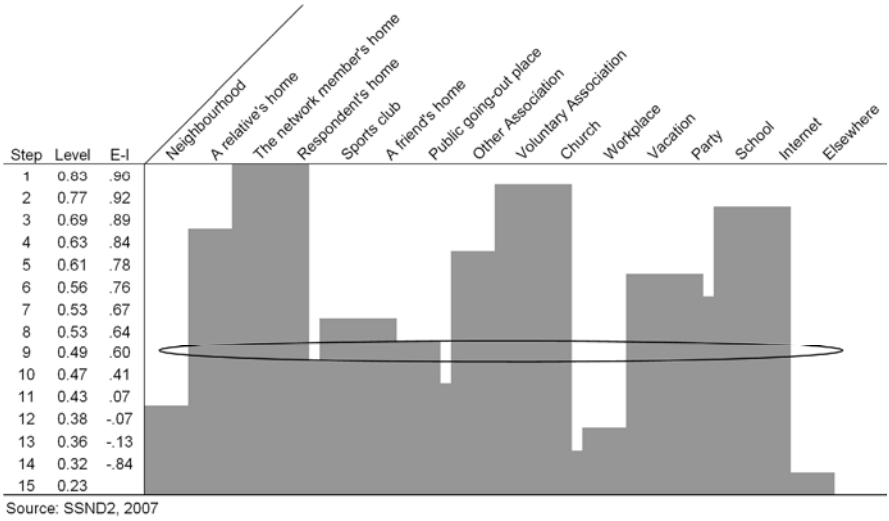
Table 3: Overlap among social contexts

Social contexts		Private				Public											
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Private	1 resp. home	3605	2790	794	368	6	153	350	126	36	87	132	56	151	262	22	45
	2 memb. home	2790	3049	696	270	5	131	340	103	28	75	129	51	115	235	18	38
	3 a relative's	794	696	967	75	0	10	8	15	2	3	7	3	15	46	7	5
	4 a friend's	368	270	75	492	0	25	22	21	8	15	15	3	17	44	1	7
Public	5 school	6	5	0	0	26	1	0	0	0	0	0	0	1	2	1	1
	6 work	153	131	10	25	1	1218	9	11	9	6	42	6	4	65	9	11
	7 neighbourhood	350	340	8	22	0	9	750	22	6	11	11	4	3	37	1	7
	8 sports club	126	103	15	21	0	11	22	202	5	2	10	2	5	11	0	1
	9 voluntary assoc.	36	28	2	8	0	9	6	5	86	6	4	12	1	8	1	1
	10 other assoc.	87	75	3	15	0	6	11	2	6	152	6	5	5	10	1	1
	11 going-out	132	129	7	15	0	42	11	10	4	6	232	1	11	15	1	2
	12 church	56	51	3	3	0	6	4	2	12	5	1	88	1	4	0	0
	13 vacation	151	115	15	17	1	4	3	5	1	5	11	1	190	23	3	2
	14 party	262	235	46	44	2	65	37	11	8	10	15	4	23	455	4	5
	15 internet	22	18	7	1	1	9	1	0	1	1	1	0	3	4	44	0
	16 elsewhere	45	38	5	7	1	11	7	1	1	1	2	0	2	5	0	371

Source: SSND2, 2007 (N = 5,894 relationships; Total number of overlaps = 8,435)

These findings confirm the first hypothesis that context overlap is more likely between a private social context and a public social context as compared to two public social contexts, but most likely between two private contexts.

The overlap structure among social contexts, as presented in Table 3, however, is distorted by the 'size' of each context. This means that two popular contexts for meeting network members show a higher overlap than two non-popular contexts, simply because in both contexts many people meet each other. To obtain the overlap structure among social contexts, controlled for 'context size', I used Bonacich' (1972) normalization method. From the resulting table (which is not presented in this paper) it is hard to see which contexts are relatively more likely to overlap than others. I therefore present results of a Johnson's (1967) hierarchical clustering analysis in Figure 1. This figure shows, for example, that 83 percent of all informal network members who meet at respondent's or the network member's home, meet at both these places, as well as that 77 percent of all network members who meet at church or at a voluntary association, meet at both these places.



Source: SSND2, 2007

Figure 1: Hierarchical clustering of social contexts

According to the 'E-I measure of cluster adequacy'[8], which is implemented in Ucinet (Borgatti et al. 2002), the solution of the ninth step of the procedure (encircled in the figure) is most tempting. It separates four clusters of contexts with relatively high numbers of overlaps: (1) a relative's home, respondent's home, and the network member's home, (2) sports clubs, a friend's home, and going-out places, (3) churches, voluntary associations, and other associations, and (4) vacations, parties, schools, and the internet.

The first cluster (respondent's home, the network member's home, and a relative's home) provides additional confirmation for the first hypothesis, i.e., that context overlaps are most likely between two private contexts. It turns out that context overlap, not only in the absolute number of overlaps, but also in relative terms most often occurs among the private contexts of one's own home, the network member's home, and a relative's home. The second cluster, however, shows that the hypothesis is not confirmed for those who meet at a friend's home, because it is relatively more likely that those who meet at a friend's home also meet at a sports club or at a public going-out place than in one of the private contexts. The third and fourth cluster show that, if the number of people who meet in each context is taken into account, overlaps are also likely among church, voluntary associations, and other types of associations, as well as among party, school, vacation, and the internet. Finally, another important finding is that the neighbourhood, but especially the workplace hardly overlaps with other contexts, neither in absolute terms, nor in relative terms. This means that the people we meet in these contexts, although they are part of our informal personal network, are most unlikely to be met in multiple contexts.

4.3 Context overlap affecting multiplexity

Table 4 presents a cross-tabulation of multiplexity (i.e., the number of exchange contents in a relationship) and the number of shared contexts per relationship. While these figures show that there is an association between these two relationship characteristics, they also confirm the argument that a distinction should be made between these two characteristics (see Feld 1981:1024-25): about forty percent of those who share more than one context do not have a multiplex relationship ([567+913]/[1,234+2,391]).

[8] This E-I index measures the ratio of the numbers of overlaps within the clusters to overlaps between clusters (cf. Borgatti et al. 2002).

Table 5 presents results from multilevel regression models on multiplexity in informal personal relationships. In all models, I controlled for respondent's age, sex, level of education, marital status, employment status, and the size of the informal personal network.

Table 4: Multiplexity and context overlap

	Social contexts per relationship			
Multiplexity of relationship	1	2	3	Total
1	1,467	567	913	2,947
2	434	311	623	1,368
3	237	208	519	964
4	108	106	230	444
5	21	41	100	162
6	0	1	6	7
Total	2,267	1,234	2,391	5,892

Source: SSND2, 2007

The first model contains two dummy-coded variables: one for those who meet each other in two different contexts, and one for those who meet in three contexts, such that those who meet in a single context make up the reference category. This model shows that sharing multiple contexts has a strong and positive effect on the multiplexity of a relationship (in accordance with the second hypothesis).

In the second model, the number of shared social contexts is split up further, indicating whether the overlap involves private or public contexts. This model shows that the positive effect of sharing multiple contexts on multiplexity predominantly applies to private contexts: in order to make a relationship multiplex, it is especially important to meet at least in one private context. The average multiplexity for those who meet each other in just one context is substantially higher if they meet in a private context instead of in a public context. Next, multiplexity increases if people share one private and one public context and even more so if they share two private contexts (both compared to sharing just one public context). Moreover, sharing two public contexts instead of one public context does not affect multiplexity. Also for those who share three contexts, it turns out to be important that at least one of the contexts they share is a private context, in order to make their relationship multiplex. So in short, these findings indicate that for a relationship to be multiplex, it is important to meet each other in multiple contexts, of which at least one has to be a private context.

Table 5: Multilevel regression models on the multiplexity of informal personal relationships[a] (with standard deviations in parentheses)

	Model 1	Model 2	Model 3
Frequency of contact			
weekly			.40 (.03) ***
daily			.46 (.04) ***
Liking each other			.20 (.02) ***
Duration of relationship			-.00 (.00)
Type of relationship			
partner			.41 (.13) **
relative, living-in			.53 (.49)
relative, not living-in			.19 (.06) ***
friend			.47 (.05) ***
co-worker			-.26 (.06) ***
neighbour			.08 (.05) ***
co-member			-.08 (.08)
Social contexts per relationship			
1 context	ref.		
2 contexts	.41 (.04) ***		
3 contexts	.61 (.03) ***		
1 context: public		ref.	ref.
1 context: private		.72 (.05) ***	.25 (.07) ***
2 contexts: both private		.84 (.05) ***	.55 (.06) ***
2 contexts: 1 private & 1 public		.65 (.08) ***	.34 (.08) ***
2 contexts: both public		.10 (.10)	.06 (.09)
3 contexts: all private		.91 (.04) ***	.63 (.06) ***
3 contexts: 2 private & 1 public		.92 (.04) ***	.56 (.06) ***
3 contexts: 1 private & 2 public		.99 (.09) ***	.57 (.10) ***
3 contexts: all public		.42 (.21) *	.23 (.19)
Constant	1.73 (.16) ***	1.43 (.16) ***	.29 (.19)
Number of relationships	5,855	5,855	5,005
Number of respondents	590	590	577
LR Chi-Squared	390.98	693.18	1307.17

Source: SSND2, 2007 (* p<0.05 ** p<0.01 *** p<0.001)

a) All models are controlled for respondent's age, sex, level of education, being married, having a paid job, and informal network size.

Finally, the third model tests whether the effect of sharing multiple social contexts on the multiplexity of a relationship is affected by other relationship characteristics. By controlling for frequency of contact with the network member, the extent to which one likes the network member, the duration of their relationship, as well as the type of relationship, this model shows that sharing multiple, and especially sharing multiple private contexts, has a robust and positive effect on multiplexity in informal personal relationships.

5 Conclusions

This study provides new insights into the structure and composition of personal networks. First, I show that the average personal network contains about ten members who altogether provide a person with personal or work-related advice from time to time, with instrumental help with odd jobs or in case one falls ill, or who give her/him company by paying a visit or by going-out together. The finding that just a small part of these informal networks can be considered as the 'core discussion network', supports the idea that people probably have "a smaller core of very close confidant ties and a much larger array of less interconnected, more geographically dispersed, more unidimensional relationships" (McPherson et al. 2006).

Second, I show that multiplex relationships are not rare nowadays, and describe which combinations of exchange contents are more likely in a relationship than other combinations of exchange contents. Especially going-out together and discussing important personal matters are often combined in a relationship, as well as giving a hand with odd jobs at home and providing help in case of illness. Relationships with those whom we discuss work-related problems with are most likely to be uniplex.

Third, I show that context overlap is not rare nowadays: people meet their informal network members in two social contexts on average. The great majority of these overlaps among contexts, however, concern overlaps with people's homes, which means that many informal network members are met at home in combination with another context. Furthermore, I show in more detail that private contexts are more likely to overlap than public contexts, and more in particular, that overlaps among two public contexts hardly ever occur. This means that 'public' and 'private' are not two worlds apart, but that 'the public world' itself is divided in separate segments with strict boundaries.

Fourth, findings in this study support the argument (amongst others made by Feld, 1981) that an analytic distinction should be made between context overlap and multiplexity in personal relationships. The empirical results support the argument that multiplexity in personal relationships is positively affected by the number of contexts people share with each other, even if I control for other relevant relationship characteristics, like contact frequency and type of relationship. More specifically, I found that sharing multiple private contexts leads to multiplexity in an informal personal relationship.

6 Discussion

More generally, this study on context overlap contributes to the research literature, as it provides an insight into the social structure of a present-day modern Western society (cf. McPherson et al. 2001). The finding that, besides the context overlaps in which people's homes are involved, other relevant social contexts in which people meet their network members hardly overlap, has some important implications.

First, a low level of context overlap implies that the argument that people live together with one set of people, work with a second set, conduct sports with a third set, and spend a night out with another set of people (as put forward by, amongst others, Fischer et al. (1977) and Coleman (1990)), is not completely correct. I show that public contexts hardly overlap each other, but also that there is still overlap among public contexts on the one hand and private contexts on the other hand.

Second, a low level of context overlap might affect the density and sociodemographic composition of the network. When each network member is met in a different context, network density is expected to be low. And because previous research on homogeneity in personal networks has shown that similarity (or homophily) in personal relationships is affected by the context in which they emerge (Marsden 1987, McPherson et al. 2001, Mollenhorst et al. 2008a, 2008b), this might also result in varied and heterogeneous networks. Another possibility, however, is that people just use a small number of contexts to meet their network members, because it is more efficient, and less costly and cumbersome. Social contexts then still hardly overlap, while the resulting networks are dense and homogeneous. In addition, previous research has revealed a certain degree of path-dependent use of social contexts for emerging personal networks (Mollenhorst et al. 2008a), but further research is needed to find out whether people continue to

meet their network members in the same contexts in which their relationships originally emerged, or that they find a new 'focus of activity' around which they can maintain their personal relationships (cf. Feld 1981).

References

Barnes, J.A. (1969). Networks and Politics. In J.C. Mitchell (Ed.), *Social Networks in Urban Situations: Analyses of Personal Relationships in Central African Towns.* Manchester: Manchester University Press.

Barnes, J.A. (1972). Social Networks, Module No. 26. In: *Anthropology*. Reading: Addison Wesley.

Bonacich, P. (1972). Technique for Analyzing Memberships. *Sociological Methodology*, 4, 176-185.

Borgatti, S.P., Everett, M.G., & Freeman, L.C. (2002). Ucinet 6 for Windows. Harvard, Analytic Technologies.

Boissevain, J. (1974). *Friends of Friends: Networks, Manipulators and Coalitions*. Oxford: Basil Blackwell.

Breiger, R.L. (1974). The Duality of Persons and Groups. *Social Forces*, 53, 181-190.

Coleman, J.S. (1990). *Foundations of Social Theory.* Cambridge, MA: Harvard University Press.

Coleman, J.S. (1993). The Rational Reconstruction of Society: 1992 Presidential Address, *American Sociological Review*, 58, 1-15.

Faust, K., & Romney, A.K. (1985). Does STRUCTURE Find Structure? A Critique of Burt's Use of Distance as a Measure of Structural Equivalence. *Social Forces*, 7, 77-103.

Feld, S.L. (1981). The Focused Organization of Social Ties. *American Journal of Sociology*, 86, 1015-1035.

Fischer, C.S. (1982). *To Dwell Among Friends: Personal Networks in Town and City*. Chicago: University of Chicago Press.

Fischer, C.S., Jackson, R.M., Stueve, C.A., Gerson, K., McCallister Jones, L., & Baldassare, M. (1977). *Networks and Places: Social Relations in the Urban Setting*. New York: Free Press.

Gluckman, M. (1962). Les Rites de Passage: In M. Gluckman (Ed.), *Essays on the Ritual of Social Relations.* Manchester: Manchester University Press.

Haines, V.A., & Hurlbert, J.S. (1992). Network Range and Health. *Journal of Health and Social Behavior*, 33, 254-66.

Ibarra, H. (1995). Race, Opportunities, and Diversity of Social Circles in Managerial Networks. *The Academy of Management Journal,* 38, 673-703.

Johnson, S.C. (1967). Hierarchical Clustering Schemes, *Psychometrika*, 32, 241-53.

Kapferer, B. (1969). Norms and the Manipulation of Relationships in a Work Context. In: J.C. Mitchell (Ed.), *Social Networks in Urban Situations: Analyses of Personal Relationships in Central African Towns*. Manchester: Manchester University Press.

Lazega, E., & Pattison, P.E. (1999). Multiplexity, Generalized Exchange and Cooperation in Organizations: A Case Study. *Social Networks*, 21, 67-90.

Marin, A. (2004). Are Respondents More Likely to List Alters with Certain Characteristics? Implications for Name Generator Data. *Social Networks*, 26, 289-307.

Marsden, P.V. (1987). Core Discussion Networks of Americans. *American Sociological Review*, 52, 122-131.

Marsden, P.V. (1990). Network Diversity, Substructures, and Opportunities for Contact. In C. Calhoun, M.W. Meyer, & W.R. Scott (Eds.), *Structures of Power and Constraint*. New York: Cambridge University Press.

McPherson, M., Smith-Lovin, L., & Cook, J.M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27, 415-44.

McPherson, M. Smith-Lovin, L., & Brashears, M.E. (2006). Social Isolation in America: Changes in Core Discussion Networks over Two Decades. *American Sociological Review*, 71, 353-375.

Mitchell, J.C. (1969). *Social Networks in Urban Situations*. Manchester: University of Manchester Press.

Mollenhorst, G., Völker, B., & Flap, H. (2008a). Social Contexts and Core Discussion Networks: Using a Choice-Constraint Approach to Study Similarity in Intimate Personal Relationships. *Social Forces*, 86 (3), 937-65.

Mollenhorst, G., Völker, B., & Flap, H. (2008b). Social Contexts and Personal Relationships: The Effect of Meeting Opportunities on Similarity for Relationships of Different Strength, *Social Networks*, 30, 60-68.

Nisbet, R. (1969). *The Quest for Community*. New York: Oxford University Press.

Pescosolido, B.A., & Rubin, B.A. (2000). The Web of Group Affiliations Revisited: Social Life, Postmodernism, and Sociology. *American Sociological Review*, 65, 52-76.

Putnam, R.D. (2000). *Bowling Alone: The Collapse and Revival of American Community*. New York: Free Press.

Skvoretz, J., & Agneessens, F. (2007). Reciprocity, Multiplexity, and Exchange. *Measures, Quality & Quantity*, 41, 341-57.

Stein, M.R. (1960). *The Eclipse of Community*. Princeton: Princeton University Press.

Verbrugge, L.M. (1979). Multiplexity in Adult Friendships, *Social Forces*, 57, 1286-1309.

Völker, B., & Flap, H. (2002). *The Survey of the Social Networks of the Dutch (SSND1): Data and Codebook*. Utrecht University/ICS.

Völker, B., Flap, H., & Mollenhorst, G. (2007). *The Survey of the Social Networks of the Dutch, Second Wave (SSND2), Data and Codebook*. Utrecht University/ICS.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge: University Press.

Wellman, B. (1979). The Community Question: The Intimate Networks of East Yorkers. *American Journal of Sociology*, 84, 1201-1231.

Wellman, B. (1999). *Networks in the Global Village*. Boulder: Westview Press.

Wellman, B., & Wortley, S. (1990). Different Strokes from Different Folks: Community Ties and Social Support. *American Journal of Sociology*, 96, 558-88.

Wheeldon, P. (1969). The Operation of Voluntary Associations and Personal Networks in the Political Processes of an Interethnic Community. In J.C. Mitchell (Ed.), *Social Networks in Urban Situations: Analyses of Personal Relationships in Central African Towns.* Manchester: Manchester University Press.

Appendix: Name generating questions

Below is the list of the 13 'name-generating' questions that were used in the SSND2 to delineate the personal networks of the respondents. To define informal personal networks, I do not make use of the role-related questions (3, 4, 5, 7, and 13), because that part of one's network is directly exogenously determined. Instead, I focus on the remaining eight questions which are marked with a star*.

1. *In case you have a problem at work, to whom do you go for advice? (5 names maximally)
2. *And the other way round? Are there people who come to for advice in case they have a problem at their work? (5 new names maximally)
3. Who are the two colleagues with whom you have to do something most frequently? (2 names maximally)
4. Who is your boss/executive? (1 name maximally)
5. People do not only have cooperative relationships, but sometimes people also bother each other. With whom do you quarrel sometimes, or who really bothers you sometimes? In short: who annoys you? (5 new names maximally)
6. *If you are doing an odd job at home and you need someone to give a hand, e.g., to carry furniture or to hold a ladder, whom do you ask for help? (5 new names maximally)
7. Who are your two most direct neighbours; that is who lives on the left, on the right, above, and/or below you? (2 names maximally)
8. *Many people visit other people in their leisure time. Whom do you pay a visit, or who visits you sometimes? (5 new names maximally)
9. *Do you know people you like going-out with, e.g., to a movie, a bar, a theatre, etc., and with whom you actually do this sometimes? (5 new names maximally)
10. *Life is usually not only about going-out and enjoying company. Everybody needs someone to talk about important matters from time to time. With whom did you discuss important personal matters during the last six months? (5 new names maximally)
11. *Whom could you ask for help in case you fall ill? Think for example of going to a grocery store or a drugstore for you. (5 new names maximally)
12. *We did already ask you for those whom you ask to help you sometimes. We would also like to know whether there are people who ask you to help them. So, who does sometimes ask you for help with an odd job at home? (5 new names maximally)
13. Now let us have a look at the list of names we have gathered so far: Is there anyone important to you, but who is not on the list yet? If so, I would like to add this person. (5 new names maximally)

Navigating and Annotating Semantically-Enabled Networks of People and Associated Objects

Sheila Kinsella, Andreas Harth, Alexander Troussov, Mikhail Sogrin, John Judge, Conor Hayes & John G. Breslin

Abstract Social spaces such as blogs, wikis and online social networking sites are enabling the formation of online communities where people are linked to each other through direct profile connections and also through the content items that they are creating, sharing and tagging. As these spaces become bigger and more distributed, more intuitive ways of navigating the associated information become necessary. The Semantic Web aims to link identifiable objects to each other and to textual strings via relationships and attributes respectively, and provides a platform for gathering diverse information from heterogeneous sources and performing operations on such linked data. In this paper, we will demonstrate how this linked semantic data can provide an enhanced view of the activity in a social network, and how the Galaxy tool described in this work can augment objects from social spaces, by highlighting related people and objects, and suggesting relevant sources of knowledge.

1 Introduction

The ability to link to other pages and objects is a key facility of the World Wide Web architecture. It has enabled every web site to become part of a global network of information. More recently, new client server applications such as wikis and blogs have made writing and linking on the Web extremely easy for the average user. The result has been the creation of vast amounts of user-generated content, often organised within online communities. Consequently, there are huge amounts of data becoming available (in real-time or near real-time), creating invit-

ing possibilities for network research, and enabling entirely new avenues for analysis.

In order to take advantage of the huge store of knowledge which is amassing online, we require new methods of navigating this data. The problem is not simply one of countering information overload, although this is certainly pertinent, but of inferring links between relevant sources of information, possibly scattered across several domains. The goal is to enable the user to move through the information space quickly and intuitively by locating relevant related people, concepts and objects at every step.

One problem is that the current link mechanism on the Web does not differentiate between different types of links and does not allow different types of relationships to be expressed. Data is presented as a set of documents and other files, interconnected by hypertext links. The concepts represented in the documents and the types of the relationships between them are not explicitly stated, and can be hard for a computer to infer. Additionally, data accumulated by one user in a particular domain cannot be easily transferred to another domain. For instance, a blogging community may be dispersed over numerous different sites and platforms, and an interest group may share photos on Flickr, bookmarks on del.icio.us, and hold conversations on a discussion forum. A single person may have several separate online accounts, and may have a different network of friends on each. Therefore, the information existing in online social spaces forms massive, intricate and generally disjoint networks of people and objects.

In short, the lack of standards for expressing semantic information in Web 1.0 has resulted in difficulties in aggregation and integration for applications and research, impairing the possibilities for data and network analysis.

Semantic Web research (Berners-Lee et al. 2001) offers the possibility of overcoming these problems by enabling the description of arbitrary objects or concepts, and the relationships between them, using shared machine-readable formats. Semantic data can be viewed as a directed graph where the nodes represent objects or concepts, and the ties represent semantic relationships. A fundamental part of the Semantic Web is the ontology, a data structure specifying the concepts that are needed to understand a domain, and the vocabulary and relationships required to enter into a discourse about it.

Representing Web data in this way allows the expression of different types of relationships between people, between people and concepts or objects, and so forth. Furthermore, these types of relationships are expressed in open formats and can be transferred and understood in the different domains or communities. For

example, the Friend-of-a-Friend (FOAF)[1] vocabulary allows for the expression of the links between people and the things they create and do. The relationships between communities of friends represented in FOAF can be processed in any program that understands the FOAF vocabulary.

There are large and detailed datasets available on the Semantic Web, containing information regarding people, their activity, and their interactions, that are amenable to social network analysis. However, there is a mismatch between two-dimensional graph theory and multi-dimensional social networks (Scott 1988). Real networks contain different types of relations, and are built around objects which connect people together. The use of semantic graphs containing heterogeneous nodes and ties, instead of traditional link-matrices, to represent information about online communities addresses this problem. For example, relation types in an online social network could include "knows" and "sent-email-to" and object types could include publications (linking authors), photographs (linking people depicted in them), and topics (linking those who have an interest in them).

Creating a graph on the Web of different types of objects linked by different types of relationships is a major step towards large-scale computational social network analysis systems that can process various kinds of relationships and objects. However, in order to fully realise the power of these new representation models, users require ways to extract knowledge from the semantic graph and to infer associations between objects that may not be explicitly linked.

In this paper we show how relevant related information can be extracted from Semantic Web data using the Galaxy tool where the output is generated by a spreading activation technique over weighted links. A related method has been applied (Amitay et al. 2004) to derive a geographical focus from a text, based on locations which are mentioned in the text, but that algorithm can operate only on a hierarchical network. Spreading activation has been applied to semantic networks for social network analysis in applications including recommender systems (Liu et al., 2006), community detection (Alani et al. 2003), and modeling trust propagation (Ziegler and Lausen 2005).

To demonstrate our technique, we gather information represented in common formats and represent the data as a semantic graph, consisting of interrelated people, objects and their associated semantic terms. This data is used as input to the Galaxy tool which provides a generic way of ontology-based network mining. We attempt to locate a set of related items within our dataset, given some text re-

[1] http://www.foaf-project.org/

ferring to a particular person or object, or to a set of people and objects. We apply our approach to two example scenarios:

- Ego-centric search, where we attempt to locate a set of nodes closely related to a focus person
- Community detection, where we locate a community centred around two focus people

Our approach makes use of the network of ties existing between people, including not only social connections, but also semantic connections via shared interests or other areas of common ground. The analysis extends further than people and objects that are closely related, to three degrees of separation and beyond.
The main contributions of this paper are as follows:

- We illustrate how a semantic data model of social spaces gives easy access to massive amounts of freely available information
- We describe how Semantic Web data can give improved insights into the activity of a social network
- We present initial results of experiments carried out on a data set extracted from the Semantic Web

2 Object-centered networks

Jyri Engeström, co-founder of the micro-blogging site Jaiku, has theorized that the longevity of social networking sites is proportional to the "object-centered sociality"[2] occurring in these networks, i.e. where people are connecting via items of interest related to their jobs, workplaces, favourite hobbies, etc. On the Web, social connections are formed through the actions of people - via the content they create together, comment on, link to, or for which they use similar annotations.

Adding annotations to items in social networks (e.g., using topic tags, geographical pinpointing, etc.) is an especially useful aid for browsing and locating both interesting items and related people with similar interests. Some popular types of content items include blog entries, videos, and bookmarks. These objects serve as the lodestone for social networks, drawing people back to check for new items and for any updates from those in their network who share their interests. On Flickr, people can look for photos categorized using an interesting

[2] http://www.zengestrom.com/blog/2005/04/why_some_social.html

"tag", or connect to photographers in a specific community of interest. On Upcoming, events are also tagged by interest, and people can connect to friends or like-minded others who are attending social or professional events in their own locality.

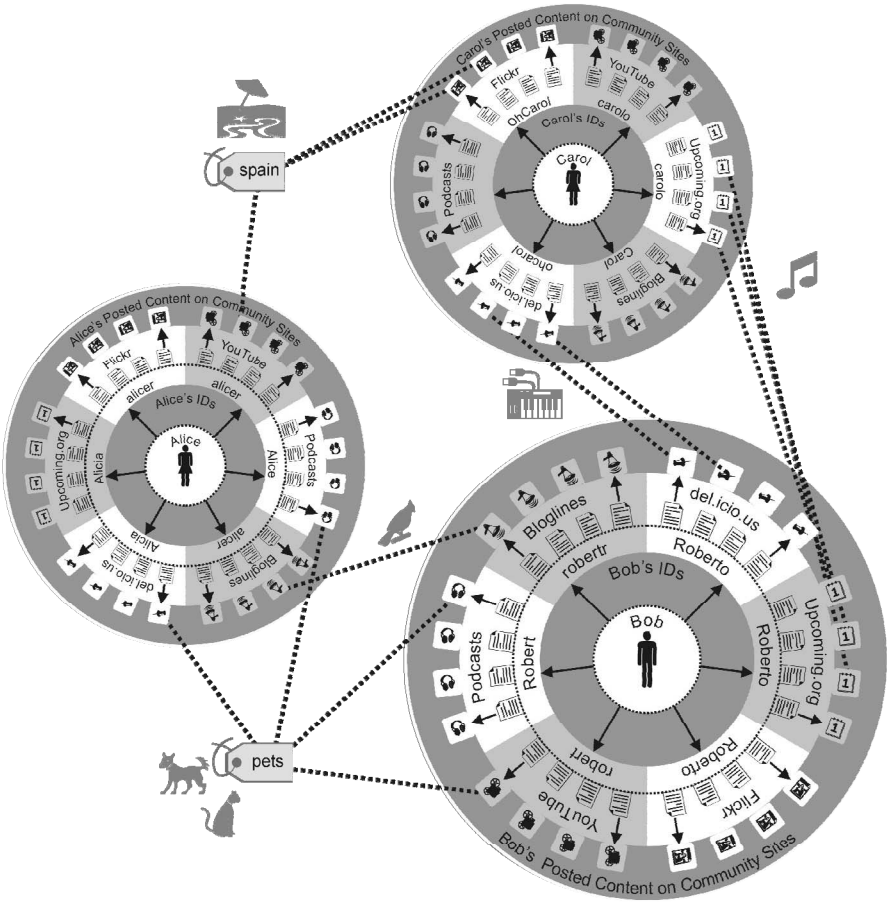


Figure 1: Object-centered social networks are formed by people (using their online accounts) and the content items they act upon

Fig. 1 is illustrative of an object-centered social network for three people, showing their various user accounts on different websites and the things that they create

and do using these accounts. Rather than being connected simply through online social network relationships (i.e. by explicitly-defined friends contacts), these people are bound together through "social objects" of common interest. For example, Bob and Carol are connected through bookmarked websites that they both have annotated on musical keyboards and also through music-related events that they are both attending. Similarly, Alice and Bob are using matching tags on media items about pets and are subscribed to the same blog on birds.

As the connections between people become intertwined with their real-world interests, it is probable that people's social networking methods will move closer towards simulating their real-life social interaction, so that people will meet others through something they have in common, and not by randomly approaching each other.

Since more interesting social networks are being formed around the connections between people and their objects of interest, and as these object-centered social networks grow bigger and more diverse, more intuitive methods of navigating the information contained in these networks have become necessary – both within and across social networking sites (e.g., a community of interest for mountaineering may consist of people and content distributed across photo-, bookmark- and event-centred social networks).

Person- and object-related data can also be gathered from various social networks and linked together using a common representation format. This linked data can provide an enhanced view of individual or community activity in a localized or distributed object-centered social network(s) ("show me all the content that Alice has acted on in the past three months").

The Semantic Web, which aims to link identifiable objects to each other and to textual strings, can be used for linking the diverse information from heterogeneous social networking sites and for performing operations on such linked data. The involvement of objects in social networks on the Semantic Web has been investigated (Kinsella et al. 2007). The Semantic Web is already being used by various efforts to augment the ways in which content can be created, reused and linked by people on social networking and media sites. These efforts include the FOAF project, ontology-enhanced wikis such as the Semantic Media Wiki, the NEPOMUK social semantic desktop[3], and the Semantically-Interlinked Online Communities (SIOC)[4] initiative. In the other direction, object-centered networks can serve as rich data sources for Semantic Web applications. Tim Berners-Lee said in a 2005

[3] http://nepomuk.semanticdesktop.org/

[4] http://sioc-project.org/

podcast, "I think we could have both Semantic Web technology supporting online communities, but at the same time also online communities can support Semantic Web data by being the sources of people voluntarily connecting things together." Users of social networking sites are already creating extensive vocabularies and annotations through "folksonomies" (collections of free-text keywords that are used to tag content items). Since the meaning of these terms is being produced through a consensus of community users, these terms are serving as the objects around which more tightly-connected social networks are centred and formed.

3 Semantic Web

The purpose of the Semantic Web is to enable the online description of arbitrary objects in such a way that software can be used to automatically combine, mine, process, and manipulate data from the Web. Machine-readable descriptions of objects and the relationships between them on the Web enable universal knowledge representation mechanisms on a global scale. For the simplest form of object identification, the same Uniform Resource Identifier is used across multiple sources to reference an object. In many people using the same URI for a particular object, the available data pieces mesh up and form a well-connected and richly-interlinked information space with structured representation features. Layered on top of the foundational URI naming mechanism are a number of other technologies to enable knowledge representation features of increasing sophistication:

- Resource Description Framework (RDF): a universal way of identifying and talking about entities, basic type system (Manola and Miller 2004)
- RDF Schema (RDFS): vocabulary with terms for describing classes and properties, subclass and subproperty relationships (Brickley and Guha 2003)
- Web Ontology Language (OWL): terms for describing classes, inverse properties, cardinality constraints; subset of first order logics (Dean and Schreiber 2004)

Information on the Semantic Web is commonly expressed using the RDF language. An RDF document is composed of a sequence of statements of the form *<subject, predicate, object>*, indicating a directed tie from the subject node to the object node, where the predicate describes the relationship between them.

On the level of RDFS, nodes represent instances of classes, and links represent instances of properties. Classes and the properties which can exist between them are defined in RDFS or OWL. The description of classes and properties form a

vocabulary that can be created or extended as required. For example, vocabularies exist to describe projects, communities, geographical information, and many other domains.

RDF uses the concept of URIs to name all sorts of objects; for example: *http://www.w3.org/People/Berners-Lee/card#i* to denote Tim Berners-Lee, *http://sws.geonames.org/2964180/* to denote the city Galway, *http://deri.ie/* to denote the research institute, and *http://purl.uniprot.orgs/uniprot/Q91474* to denote the protein SHNF1. Objects identified via URIs typically have one or many associated types e.g. *http://xmlns.com/foaf/0.1/Person* or *http://swrc.ontoware.org/ontology#FullProfessor*. The relationships between objects are denoted using URIs, such as the instance-to-type relation *rdf:type*. Namespace prefixes (such as *rdf:*), which indicate the schema to which classes and properties belong, can be used to abbreviate URIs.

In Semantic Web research, the standard way to infer knowledge from a semantic graph is to use an inference engine based on a logic framework such as the OWL to allow logic reasoning on the Web. However, inferring general relationships from graphs can be achieved using techniques other than logic, as we demonstrate in this paper with Galaxy.

4 Dataset

We analyse a dataset consisting of social network information focused around the Semantic Web community. Our model includes people and various related entities. The data under analysis is part of a web crawl of RDF data that was carried out during June/July 2007 using MultiCrawler (Harth et al. 2006). The initial dataset originates from approximately 85,000 sources and consists of over 35 million statements. Object consolidation (Hogan et al. 2007) was performed in order to merge identifiers of equivalent instances occurring across different sources. From the original crawl, we extracted a smaller sub-graph for analysis. The sub-graph is based around the URIs of four people in the Semantic Web community: Tim Berners-Lee, Dan Brickley, Andreas Harth and Tim Finin. We used YARS2 (Harth et al. 2007) to extract all people connected by a path of three or less ties to any of the root nodes, via *foaf:knows* relations. We also included any other nodes connected to these people. The resulting dataset consists of a vast amount of information in many different vocabularies, totalling over 1.2 million statements.

The current version of Galaxy is an early prototype which takes input data expressed in an XML format. However it is planned that RDF support will be

available in the near future. We developed a program to extract specific information from RDF and map it to the required format. For this initial work, we include only a small set of relation types, but it would be possible to extract a much broader range of data. The information we extract is a subset of three vocabularies. Most of the dataset is described using the Friend of a Friend vocabulary (shorthand:*foaf*), which enables the description of people and their relationships with other resources. It also enables the expression of other information relating to a person, such as contact details, as well as publications and other items they have created. Anyone can create their own FOAF file describing themselves and their social network, and social network services can also automatically generate FOAF files for their users, as some, for example LiveJournal, already do. The demand for open, common standards like FOAF is evident from the recent interest in DataPortability[5], an effort by providers of social software, such as Google, Facebook and LinkedIn, to enable users to control and share data across different websites. The information from multiple FOAF files can easily be combined to obtain a higher-level view of the network. We also include some data expressed using the RDF Schema (shorthand:*rdfs*) and Dublin Core (shorthand:*dc*)[6], both of which include properties commonly used to specify the names of resources. There are two main steps to the conversion process - extraction of nodes and ties, and extraction of text labels.

Table 1: FOAF predicates which were extracted and the relation type to which they were mapped

Predicate (foaf:)	Relation type
knows	knows
interest	hasInterest
currentProject, pastProject	hasProject
workInfoHomepage, workplaceHomepage	hasWorkplace
schoolHomepage	hasSchool
made	isMakerOf
maker	madeBy

We derive information from RDF statements based on predicates. All extracted nodes and ties are assigned a type. For instance, all object nodes which occur with the predicate *foaf:interest* are mapped to type ‘interest’. The predicates which we

[5] http://dataportability.org/

[6] http://dublincore.org/

extracted are shown in Table 1, along with the relation type each predicate was mapped to.

Fig. 2 shows the node types which exist in our data model, and the relation types which connect them together. The predicate *foaf:maker* is the inverse of the predicate *foaf:made*. Therefore the corresponding relation type "isMakerOf" is considered to be the inverse of the relation type "madeBy"; in other words, they represent the same relationship, but in opposite directions. None of the other RDF predicates in the data we extracted have an inverse.

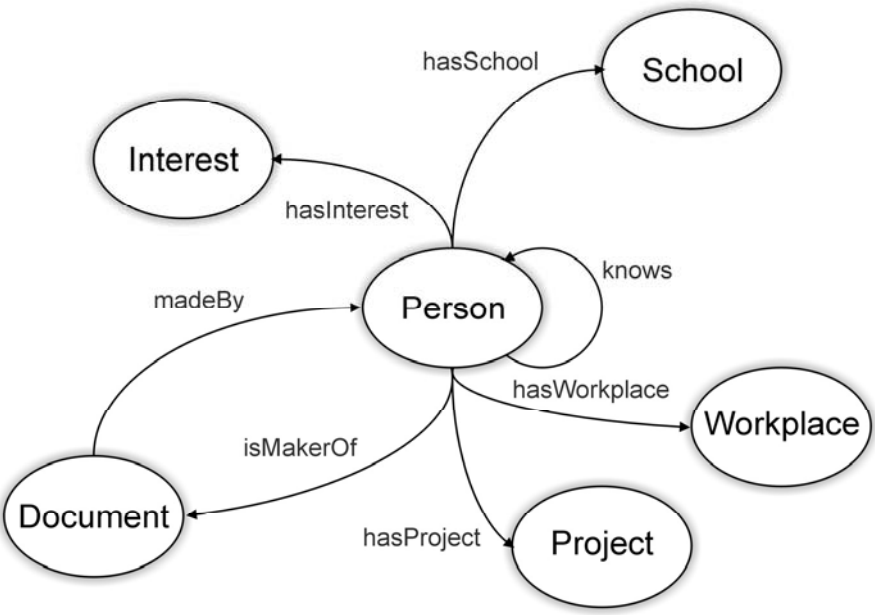


Figure 2: Node and relation types in the data model

We also extract labels for nodes, so that textual references to a particular node will be recognised. For each node type, we made a list of the predicates which indicate that the object node is a name for the subject node. For example, where the subject node is of type Person, this list includes predicates such as *foaf:nick*. Table 2 shows for each node type the predicates which we assume to indicate names. Some nodes may have many different labels. If a node has no name specified, we use the URI of the node as a label.

Our dataset contains 16468 entities and 25028 relationships. Most of the entities are people. The composition of the dataset is shown in Table 3.

Table 2: Node types and the predicates which indicate names

Type	Names
Person	foaf:nick, foaf:name, foaf:firstName, foaf:givenname, foaf:family_name, foaf:surname
Interest	dc:title, dc:subject, rdfs:label
Project	dc:title, dc:subject, rdfs:label
Workplace	dc:title, dc:subject, rdfs:label
School	dc:title, dc:subject, rdfs:label
Document	dc:title, dc:subject, rdfs:label

Table 3 Frequencies of node types in the network

Node Type	Instances	Node Type	Instances
Person	11314(68.7%)	Workplace	443(2.7%)
Interest	2228(13.5%)	Project	339(2.1%)
Document	1956(11.9%)	School	188(1.1%)

5 Galaxy

Galaxy is an ontological network miner designed by the IBM LanguageWare Team[7] for application to tasks in social semantic computing. The Galaxy tool uses a spreading activation algorithm to perform clustering on semantic networks. Instead of the traditional method of hard clustering, which partitions a graph into different groups, Galaxy performs soft clustering, which involves taking a sub-graph based around a set of input nodes, and finding the focus of this sub-graph. The method can be applied to social networks, company organisation charts, or any other set of graph-structured data. Initially, an ontological network of concepts and related terms must be generated based on data provided by the user. Galaxy can then process documents, and identify their main concepts, based on the ontological information. The two main steps to this process are the mapping of terms to concepts, and the location of the main concepts.

[7] http://www.alphaworks.ibm.com/tech/lrw

The Galaxy tool takes a piece of text as input, and then maps terms in the text to concepts in the ontological network. If necessary, the topology of the graph is used in disambiguating terms in the document. The concepts which are identified as corresponding to terms in the text act as input nodes for the spreading activation algorithm. The result of the algorithm is a set of focus nodes, which can be interpreted as those nodes which are most central in the sub-graph based around the input set.

Cognitive psychology and artificial intelligence research model reasoning and memory as processes on neural networks. These networks of neurons and the patterns in which they fire simulate certain aspects of the human brain. There are many different algorithms and implementations which model these processes, one of which, spreading activation (Anderson 1983), is implemented in Galaxy.

In general, the spreading activation algorithm proceeds as follows:

1. Initial activation is set to one or several nodes in the network (e.g. with value 1.0). This initial activation may represent items of interest, context of a document, user profile, etc., and is analogous to sources of light.
2. Activation is spread to neighbouring nodes, but the activation value is normally less then the value of a source. For this, an activation decay parameter is introduced, usually in the range [0…1]. As the activation spreads through the network, different link types may have associated different decay values allowing for different effects like a lower rate of decay through "preferred" links.
3. If activation is spread from a node with many links, those neighbouring nodes will get even less activation to simulate a situation that many similar items get less attention when compared to one unique item.
4. However, if there are multiple paths in the network to some node, its activation will be sum of activations from its inputs. And therefore, it may get activation value even higher than the source.
5. After all activation values are calculated, they are ranked and nodes with higher activation represent important or interesting items or concepts.

Fig. 3 shows how the algorithm finds the focus in a simple linear graph by propagating light of intensity 1 from the nodes at opposite ends of the graph. If the activation is allowed to propagate outwards from the starting points a central node is "illuminated" by both nodes meaning that the level of light is greatest at that point so it is chosen as the focus.

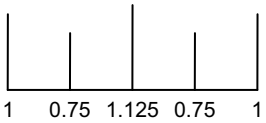


Figure 3: Illustration of the spreading activation algorithm

Galaxy can be used with any kind of graph or tree, and allows for both directed and undirected ties. Various parameters can be tuned to alter the behaviour of the algorithm. This allows a domain-expert to stipulate the properties of a semantic graph that are most important for a particular task. For example, in a graph with different types of relations, some may be considered more relevant than others, depending on the application. The Galaxy tool allows for relation types to be weighted in order to reflect the relative significance of different relationships.

Galaxy can be customised to a range of tasks. Possible application areas include expert-finding, metadata creation, and community detection.

6 Results

In the following we present the results of some sample queries for the semantic graph described in Sect. 4. In each case, we provide Galaxy with a short piece of text, and it uses the topology of the semantic network to extract the most strongly related nodes, based on terms mentioned in the text. Each instance is represented by a URI corresponding to that resource, but here we display human-readable text names. Our queries involve three people: John Breslin, Tim Berners-Lee and Andreas Harth. Firstly, we perform queries for each of these individuals in order to obtain an ego-centric view of their network. Secondly, we perform queries involving pairs of individuals as a means of detecting the community to which they belong.

The objective of the ego-centric queries is to derive an overview of the most relevant available content relating to a particular person. The results for Query 1, “John Breslin”, are shown in Table 4. For this query, Galaxy identifies the people John Breslin and Hannes Gassert, as well as several entities directly related to John Breslin and two entities related to his direct connection Hannes Gassert (Semantic Web at del.icio.us and mediagonal).

Table 4: Results for ego-centric search Query 1: “John Breslin”

Type	Instances
Person	John Breslin
	Hannes Gassert
Interest	Semantic Web at del.icio.us
	Semantic Web
	RDF
Document	John Breslin's blog
Workplace	Semantic Web Cluster, DERI
	DERI
	DERI Galway
	Líon Project, DERI
	Mediagonal
School	National University of Ireland, Galway

The results for Query 2, “Tim Berners-Lee”, are given in Table 5. Galaxy locates the appropriate person and additionally one interest and several documents. uery 3 for “Andreas Harth” locates the person Andreas Harth, one interest and two projects, as shown in Table 6.

Table 5: Results for ego-centric search Query 2: “Tim Berners-Lee”

Type	Instances
Person	Tim Berners-Lee
Interest	Semantic Web
Document	FOAF Document for Tim-Berners Lee
	Tim Berners-Lee's blog
	N3Logic : A Logic For the Web
	Creating a Policy-Aware Web: Discretionary, Rule-Based Access for the World Wide Web
	Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web
	Semantic Web Boot Camp 2007 data

Table 6: Results for ego-centric search Query 3: “Andreas Harth”

Type	Instances
Person	Andreas Harth
Interest	Knowledge Representation
Project	YARS
	SWSE

The objects retrieved by these queries are those which are found to be most relevant to the focus person; not all related entities are shown. The data on which the results are based originates not just from the FOAF files of the individuals involved, but also from other documents which contain references to these people. Results like these could be useful to someone who has come across a reference to these people on the Web and is interested in finding out more related information.

We also experimented with using Galaxy to identify a community, starting with multiple individuals within that community. We chose two queries, each mentioning two people: "John Breslin, Tim Berners-Lee" and "John Breslin, Andreas Harth". The results of these queries are shown in Table 7.

Table 7 Results for community detection Queries 4 and 5

Query	Query 4: “John Breslin, Tim Berners-Lee”	Query 5: "John Breslin, Andreas Harth"
Results	John Breslin	John Breslin
	Tim Berners-Lee	Andreas Harth
	Dan Brickley	Hannes Gassert
	Eric Miller	Aidan Hogan
	James Hendler	Matteo Magni
	Henry Story	Fergal Monaghan
	Charles McCathieNevile	Sheila Kinsella
	-	Siegfried Handschuh
	-	Axel Polleres
	-	Knud Möller

The subjects of our first community detection query, John Breslin and Tim Berners-Lee, are both involved in Semantic Web research. However they are not directly connected to each other. The results show that Galaxy has identified a set of individuals who are located around the two subjects in our query, resulting in a broad view of the Semantic Web community. These people were not identified as

relevant to either of our initial separate queries for John Breslin and Tim Berners-Lee, however when we take the two people together they are found to be important. This is because the activation spreading from both of these nodes overlaps at the nodes in between and raises their rank in the results. These results are based on data aggregated from Tim Berners-Lee's FOAF file, John Breslin's FOAF file, and other documents. This overview of the network is not possible without considering information from multiple sources in our dataset.

The second community detection query involves John Breslin and Andreas Harth. In this query the two people are again Semantic Web researchers, however in this case they work together within the same research institute. The second query therefore has a much narrower focus than the first. All of the people identified by Galaxy for the query "John Breslin, Andreas Harth" are members or former members of the Digital Enterprise Research Institute, and are closely connected to one or both subjects of the query. Most of them were not identified in Queries 1 or 3, because the connection to the focus node was not rated as strongly as, for example, documents authored by the focus node. However in the community detection queries there are now two focus nodes, and the people in the results set are included because they are related to both, which increases the activation of these nodes. As for the previous query, the results are enabled by the aggregation of social networks expressed in multiple interconnected FOAF files.

Although all of the queries given above are very simple, longer text documents can be analysed with Galaxy, for example e-mails and blog posts.

7 Discussion

The examples we have shown in this paper indicate that mining the graph of Semantic Web data using a spreading activation approach allows for the discovery of new relationships between nodes. Evaluating the results returned by Galaxy a more objective way will be a difficult task. This is due to a number of factors. The most common evaluation approaches for recommender-type systems are performed offline using techniques from machine learning and information retrieval such as cross validation and measures of recall/precision (Hayes et al. 2002). In order to conduct such an analysis we require a data set (an ontology), a number of queries, and relevance judgements for those queries on the data set. As a result of difficulties arising from these requirements we have been unable to provide an extensive qualitative analysis here.

Queries are relatively easy to create using use cases and scenarios, however, it should be noted that depending on the user or the task the same query might anticipate different results.

The data available to us is useful for proof of concept testing, but contains a lot of noise and much manual intervention was required to make the subset used in these experiments usable. Due to the novelty of our component's implementation there exists no external standard corpus or dataset (to our knowledge) which is suitable for evaluating this kind of functionality and the cost of manually creating a sufficiently large dataset is prohibitively high.

Relevance judgements for queries require a lot of manual work and investigation, are very subjective depending on who decides what is relevant, and it is very difficult to say if the process has been exhaustive on datasets large enough to be considered suitable for meaningful evaluations.

We see our work as a first step towards using rich web data to gain improved and timely insight into the formation and evolution of social networks.

8 Conclusions

This paper shows how to aggregate and integrate social network information from multiple online sources. We have demonstrated that Semantic Web technologies allow for the collection of real-world data under liberal licenses at an unprecedented scale and at a low cost. We illustrated the benefit of Semantic Web data for social network analysis using the Galaxy tool, which generates a set of related items by a spreading activation technique over weighted ties. We began with an outline of object-centered networks, and described how a semantic data model of social spaces can give an improved insight into the activity of a social network. We then explained the capabilities of the Galaxy tool in ontology-based mining of social semantic networks, and demonstrated how it can provide an enhanced view of networked data. Finally, we presented initial results of experiments carried out on a data set extracted from the Semantic Web, which make use of the network of ties existing between people, including not only social connections, but also semantic connections via shared interests or other areas of common ground. The analysis extends further than people and objects that are closely related, to three degrees of separation and beyond. There are challenges in evaluating the output of systems using web data, and the usage of personal details, even those that are publicly accessible, may create privacy concerns. However we believe that the applications and types of analysis made possible by the free availability of massive

amounts of social information will also give social network researchers a chance to work with huge amounts of real-world data and potentially gain insights into how social networks, both online and offline, form and evolve.

References

Alani, H., Dasmahapatra, S., O'Hara, K., Shadbolt, N. (2003). Identifying Communities of Practice through Ontology Network Analysis. *IEEE Intelligent Systems*, 18, 18-25.

Amitay, E., Har'El, N., Sivan, R., Soffer, A. (2004). Web-a-where: geotagging web content. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. NY, USA.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22, 261-295.

Berners-Lee, T., Hendler, J., Lassila, O. (2001). The semantic web, *Scientific American*, 284, 28-37.

Brickley, D., Guha, R. V. (2003). RDF Vocabulary Description Language 1.0: RDF Schema. *W3C Working Draft*.

Dean, M., Schreiber, G. (2004). *OWL Web Ontology Language Reference.*

Harth, A., Umbrich, J., Decker, S. (2006). MultiCrawler: A Pipelined Architecture for Crawling and Indexing Semantic Web Data. *Proceedings of the 5th International Semantic Web Conference*. Athens, GA, USA.

Harth, A., Umbrich, J., Hogan, A., Decker, S. (2007). YARS2: A Federated Repository for Searching and Querying Graph-Structured Data. *Proceedings of the 6th International Semantic Web Conference*. Busan, Korea.

Hayes, C., Massa, P., Avesani, P., Cunningham, P. (2002). An on-line evaluation framework for recommender systems. *Proceedings of the Workshop on Recommendation and Personalization in eCommerce at the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems*. Malaga, Spain.

Hogan, A., Harth, A., Decker, S. (2007). Performing Object Consolidation on the Semantic Web Data Graph. *Proceedings of the I3: Identity, Identifiers, Identification Workshop at the 16th International World Wide Web Conference*. Banff, Alberta, Canada.

Kinsella, S., Harth, A., Breslin, J. G. (2007). Network Analysis of Semantic Connections in Heterogeneous Social Spaces. *Proceedings of the UK Social Network Conference*. London, United Kingdom.

Liu, H., Maes, P., Davenport, G. (2006). Unraveling the Taste Fabric of Social Networks. *International Journal on Semantic Web and Information Systems*, 2, 42-71.

Manola, F., Miller, E. (2004). *RDF Primer*.

Scott, J. (1988). Trend Report: Social Network Analysis. *Sociology*, 22, 109-2

Ziegler, C. N., Lausen, G. (2005). Propagation Models for Trust and Distrust in Social Networks. *Information Systems Frontiers*, 7, 337-358.

The Flow of Information in Evolving Social Groups

Wolfgang Sodeur & Volker G. Täube

Abstract In the field of Social Network Analysis various concepts for the identification of subgroups in relational empirical networks as well as different theoretical aspects linked to different kinds of subgroups have received a lot of attention over the last years. In this context the problem of the consequences that arise from the choice for a certain grouping procedure for the object under investigation (e.g. the process of the spread of information) are rarely addressed. Hence, this is not very surprising since process data, that would allow a respective examination, are rather scarce. The data presented in this paper give this seldom opportunity and permit (at least rudimentary) the confrontation of a chosen grouping and the real course of the process. Such an examination is as well important for the decision on an adequate grouping procedure as it is for the identification of criteria which might be used for the description of the information flow within and between groups. Under some circumstances, a large part of the course of the information process may already be explained by a certain grouping.

1 Introduction

In the winter of 1978/79 the developing relations over the first 9 weeks amongst about 200 freshmen in the branch of economics at a German University were investigated (Projektgruppe Studienanfänger, 1982). Diary entries over the first 5 weeks as well as sociometric questions on expressive (socio-emotional) relations over the first nine weeks were equally part of the survey program as were questions about more instrumental contacts like the exchange of information on the structuring of the studies, talks about books and the acquisition of books that were relevant for the courses, etc.. With regard to the last point an experiment was con-

ducted in which 28 randomly chosen students were informed about the possibility to purchase a study-related book at a special price with the project group. It was than traced which students in the following weeks actually acquired the book. The complete data set has been documented and is henceforth freely available from the following links:

- http://soziologie.uni-duisburg.de/~hummell/freshmen
- http://www.mpifg.de/~lk/networkdata/freshmen/Freshmen_data.html

The focus of this paper is on a couple of special aspects from the above mentioned survey. In a first step we look for a social grouping among students which may be derived relatively easily and will have explanatory power with regard to the communication process in the whole population over time. In a second step we analyze the available data to partly prove whether the defined grouping indeed shows the assumed consequences.

For the identification of social groupings different procedures have been described within the field of social network analysis, e.g. identification of cliques, clans or components (Wasserman and Faust, 1994, chapter 7). The different procedures produce different groupings to which certain theoretical consequences for the cohesion of a group and the processes inside it are assigned. In the case at hand we search for the identification of groupings which are of importance for the exchange of information on a study related book amongst freshmen. Whilst there exist a vast literature on theoretical discussions about possible characteristics of formal grouping procedures testing by means of empirical criteria for the actual course of the processes within and between delimited groups is comparatively rare.

The initially mentioned data set offers a couple of possibilities for testing on the consequences of different grouping procedures for the documented processes. Comparative inquiries which lend themselves to be conducted by means of this complex data set will be postponed to later analysis. As a first introduction we will instead try to trace if a given grouping procedure chosen by theoretical reasoning on the basis of certain figurations of components leads to an explanation on how the communication process about book purchases amongst students has happened. In section 2 the chosen grouping procedure will be discussed. At the end of this chapter the focus is on some problems on the border line of substantial interests and methodological realization. This discussion yields mainly at the identification of an appropriate standardization procedure for the relative frequencies of relations that would allow for a more comprehensive handling of the data. In section 3 we describe how criteria data may be generated based on the information on actual

book purchases over 6 weeks and statements of the students about book related talks with other fellow students. Against the background of the previously defined groupings such data may give a clue on the course of the communication processes. In the final chapter 4 we merge the data from chapter 2 and 3 in order to check for the contribution of the chosen grouping of students with regard to the explanation of the assumed communication processes.

2 Groupings with a particular importance for information exchange

With regard to the need for relational data that allows for an appropriate grouping of students it would have been obvious to rely on the answers on the sociometric questions referring to instrumental aspects of the studies or even on book purchases. However, with regard to our previously formulated aim and the criteria variables this would have lead us close to a circular reasoning since we want to gain insights on how the results of a certain grouping procedure contributes to the explanation of information diffusion.

Therefore, the search for socially relevant groupings follows in a first step more general criteria for "small" social groups. Contrary, the inquiry of the consequences of such groupings for the information exchange (book purchase) will be related to data that are relatively independent from the first mentioned complex of data (chapter 3). For the identification of the evolving groups we refer to responses on three sociometric questions of expressive content:

- With which fellow students would you preferably go out on an evening? (Question No. 2)
- If you are preparing a collective homework with which fellow students would you preferably work together (besides specific competencies on a given topic)? (Question No. 4)
- Are there any fellow students with whom you would go for a holiday trip? (Question No. 5)

We interpret an immediate directed relation between two persons i and j (i→j) as existing if i mentioned person j on at least two out of the three questions. The respective relational nets were calculated for the survey weeks 2, 3, 4, 5, 7 and 9. These nets were compiled to a single one where a relation i→j was coded as being existent if it has been reported in at least two of the six weeks. Both limitations aim at integrating only relations with a certain stability into the analysis.

Against the background of the existing data and due to our theoretical interest, the sociometric concepts of a (1-) clique (each actor reaches every other actor within one step) respectively the alleviated form of a n-clan are no satisfactory solutions for the identification of groupings: the identification of n-clans leads from the small size of existing groupings amongst the freshmen to numerous groups. An explanation of diffusion processes in the whole population would therefore concentrate on the question how information flows – e.g. via brokers (see Täube, 2004) – process while transgressing delimited n-clans.

For the examination of components the converse problem becomes an issue; such groupings connect most members of the whole population and the comparison of information flows within and between groups becomes impossible. As a solution to this problem, one often refers to a compromise between the two grouping procedures while trying to identify a classification which sorts persons into the same groups if they are relatively similar with regard to certain relational properties. In the field of social network analysis the existing or absent pair-wise relations with other members of the population are considered important. Especially with directed relations the focus is more differentiated on outgoing and incoming relations to or from all other persons. However, due to our intention to explain aspects of communication processes that often go beyond dyadic exchanges the attention will not focus exclusively on direct (one-step) relations but include indirect ones as well: in case of a missing immediate directional tie i→j between each two students the possibly existing indirect connections within 2, 3 or more steps are allowed for. Thus, the network will be described for each directed pair of two students i→j by a value which denominates the number of geodesic steps between the two persons. Special cases are the value "8" with the signification "eight or more steps" and "9" indicating "no connection". For the sake of brevity we abstain from a discussion on the pros and cons of this kind of codification. We will abstain especially from a discussion of the problems that arise from interpreting figures related to connectivity (number of steps, maximum number of steps (8), no connection (9)) as being measurements on an interval scale level. Obvious is, however, that direct relations (one-step) should obtain a relatively higher weight than indirect relations, hence, interpretation of the data as being on an ordinal level should not be problematic. Other codification with alternative evaluations of direct and indirect connections of different length may lead to different results. The rows in table 1 show for example the (direct and indirect) relations with the first 20 actors of the network used to describe the first two students (fictitious data):

Due to the chosen characteristics (direct and indirect relations) each of the 182 students is described by a vector of 364 records: 182 of these records refer to per-

sons reachable (named) directly or indirectly by ego while the other 182 records refer to the reported direct or indirect connections of all others to him.

Table 1: Fictitious n-step connections between students

Ego	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	9	9	9	9	3	9	1	9	9	9	1	2	4	9	9	9	2	1	2
2	1	1	9	9	9	3	9	1	9	9	9	1	2	4	9	9	9	2	1	2

Comments:

- The numbers in the matrix show in how many steps the reporting person (here "001" in line 1 and "002" in line 2) is connected with the person in the column (here 1, 2, ...20). The "3" in column 6 (both lines) means: actor "001" (like actor "002") names a person x1, which names person x2, which finally names person "6": 001→x1→x2→6.
- A "9" depicts the case where no direct or indirect connection between the two persons (line→column) is present. Indirect connections with a length of more than 8 steps are therefore coded as "8".
- The "reflexive" connection of a person with herself ("1" in the diagonal) is given by definition. This assures the "conformity" of this characteristic in the case of mutual mentioning. For example: actor 002 mentions actor 1 and actor 001 "mentions himself" (first column); the outcome is "conformity" of both actors. Actor 001 doesn't mention person 2 and is also not connected indirectly with person 2 in this direction; therefore, the results on this property show no conformity - or a great difference - between both actors.

The classification was calculated by means of the K-MEANS-algorithm in a version from Helmuth Späth (1975, p. 73). Classifications with 5, 6, 7, ...10 classes were determined. In order to avoid local optima the procedure was carried out 10 times with randomly chosen initial arrangements. Finally, the best classification was taken into consideration respectively. We abstain here again from a detailed discussion on the justification of the number of classes chosen and confine all further analysis on a classification with 5 classes.

The heading of table 2 shows the distribution of the 182 freshmen on the 5 classes based on the number of *direct and indirect* expressive ties in both directions amongst them. Part a) of table 2 shows the absolute numbers of direct relations within and between the 5 classes. These figures can of course not be compared with each other since they depend on the respective number of members of the classes. A common procedure for the calculation of the relational density standardizes the absolute frequencies by using the maximum possible number of relations per class as a basis: within a class n (n=1, 2, ...5) of size kn are kn*(kn-1), between two classes of size kn1 and kn2 are kn1*kn2 relations possible. For example, class 1 comprises 86 knots, hence the total number of possible relations is

N * (N-1) = 86 * 85 = 7310 links. 74 links of class 1 point into itself, hence the density in per cent is 74*100 / 7310 = 1.012%. For the density of relations between class 1 and class 2 we calculate again the total number of possible relations which are due to the class sizes (N1 * N2) 86 * 19 = 1634. A total of 6 links point from class 1 into class 2, hence the density is 600 / 1634 = 0.367%; the corresponding densities (relative frequencies) are shown in part b) of table 2.

Table 2: Classification in 5 classes corresponding to the direct and indirect expressive relations in weeks 2-9 and direct relations within/between classes

Frequencies of nodes per class (total)

Class	1	2	3	4	5	Sum
Frequency	86	19	25	24	28	182

Total size of network (isolates included): 182
Total number of links: 338
Number of links among knots with class assignment: 338

a) Absolute frequencies of links within/between classes

	Class 1	Class 2	Class 3	Class 4	Class 5	Sum
Class 1	74	6	0	4	5	89
Class 2	1	37	0	0	2	40
Class 3	11	3	41	1	23	79
Class 4	2	1	0	51	2	56
Class 5	2	1	0	0	71	74
Sum	90	48	41	56	103	338

b) Relative frequencies (in % of possible links)

	Class 1	Class 2	Class 3	Class 4	Class 5	Out (Between Classes)
Class 1	1.012	0.367	0.000	0.194	0.208	0.182
Class 2	0.061	10.819	0.000	0.000	0.376	0.097
Class 3	0.512	0.632	6.833	0.167	3.286	0.968
Class 4	0.097	0.219	0.000	9.239	0.298	0.132
Class 5	0.083	0.188	0.000	0.000	9.392	0.070
In (Between Classes)	0.194	0.355	0.000	0.132	0.742	

The density of relations within a class may be found in the diagonal in part b) of table 2. Apart from the diagonal the value of the row i and the column j refers to the density of the relations of class i to class j. In addition, part b) of Table 2 shows in the last column on the right ("out") and in the footing of the table ("in") – complementary to the diagonal – the accumulated densities between the respective class and all other classes; i.e. the value of the last cell in the first line (value 0.182) shows the relational density of class 1 to all other classes (2-5).

As the diagonal in the frequency table shows, class 1 is internally only very weakly connected (1.012 percent of the highest possible density) and has also not many external relations. The similarity of its members results less from their tight connectedness but more from their common isolation. Contrary, classes 2, 4 and 5 are marked by a relatively higher share of internal relations (10.819, 9.239 and 9.392 percent) and comparatively fewer external relations. These classes may preferably be described as "more densely knitted" groupings. Finally, group 3 has also a high internal density (6.833 percent) but in addition also a fairly high number of relations to class 5 (3.286 percent). Furthermore, the stronger outward orientation of class 3 is not reciprocated by class 5. It seems that in general other classes are extremely little oriented towards class 3 (see column 3 of the relative frequencies).

Against the background of these results on the classification of expressive relations we would await especially within classes 2, 4, and 5 a stronger concentration on instrumental communication contacts. With regard to class 1, however, we assume only processes of low intensity within as well as with regard to exchange relations with other classes due to its relative isolation. For class 3 the judgment is more uncertain because of the asymmetric character of the relations: in fact, the relational density within this class is comparably high but the same holds for its outgoing relations to class 5, showing a higher density than the relations between any other classes.

At this point some remarks about standardization procedures are to be made. As mentioned before we are interested in the consequences of certain classifications or groupings of actors that are based on expressive relationships for the explanation of the communicative behavior concerning instrumental issues. In a first step we derived such a grouping and looked closer to its "meaning" while comparing the different frequencies of expressive relationships within and between groups (classes). In order to avoid biases introduced by group size we used the common procedure of standardizing the absolute frequencies by means of the respective maximal possible number of links.

In a very similar way we will compare the frequency distributions of instrumental relationships within and between different groups at the end of this paper (chapter 4). Because the usual standardization procedure leads in this context to somehow problematic results we will discuss some alternatives that will be applied throughout the analysis. We start by taking table 2 as an example.

When we compared above the within-density of the 5 classes, we intended to interpret this result as a consequence of different and class specific orientations among the respective class members that result in different "class properties". A special case of "orientation among class members" may be related to the status of isolated persons, i.e. students who did not name any other students and who were not named directly by other students. In a class with many isolates, the density will - other things being equal - be relatively low and vice versa. Indeed, the relative frequencies of isolates differ between the 5 classes. So we are able to describe this special factor separately to avoid any mixture with effects of other "orientations" among students. The first part of table 3 describes the distribution of the remaining subpopulation of 155 non-isolated students on the 5 classes.

With very similar intentions we are looking for modifications of standardization procedures. In a first step the absolute frequencies were standardized by the maximum possible number of links within or between classes (see table 2): This commonly used procedure will exclude "mechanically" the bias due to different class sizes. Substantially however, this procedure will possibly induce another bias in the opposite direction: If the contact capacities of students are assumed to be limited, the larger a class (group) will be the less probable are all of its members to be chosen by others. Consequently, the limited contact capacity of class members should be taken into account, either generally by means of the average density of links in the whole population or more specific by means of the distribution of either outdegrees or indegrees, or even of both over all classes.[1] In this paper we will take a radical perspective and will control for both, class specific means of indegrees and outdegrees. In addition to the exclusion of the isolates, we now exclude the special effects of different indegrees and outdegrees between classes from the analysis. Comparisons between the remaining relative densities within or between special classes are now reduced to the question: however the capacities of incoming or outgoing contacts of members of different classes may differ –

Are their contacts equally (randomly) distributed upon members of all classes or do they concentrate on special classes?

[1] In a separate paper (forthcoming) the authors discuss several alternate standardization procedures and explicate some assumptions which could guide the actual choice.

Given this strict perspective we concentrate on the problem whether a given classification of the student population by means of their expressive relationships induces a concentration of within-class contacts. Do classes differ with respect to their within-class-contacts? Can we identify special pairs of classes with extraordinary low/high contact rates between them?

In order to clarify these questions we start off by turning again to the distribution of links within / between classes shown in part a) of table 2, presenting shortly the standardization procedure that will be applied throughout the following analysis:

Part a) of table 3 contains the expected values of contacts within / between classes and resulting chi square and Cramers V figures, while part b) presents the empirical values as percentages of the expected values: e.g. class 1 has 74 links within itself and an expected value of 23.698, hence the relative frequency which might be interpreted as the relative concentration of links within this class is (74*100) / 23.698 = 312.26%.

Table 3: Direct relations within/between classes standardized by expected frequencies of links within/between classes i, j

Frequencies of nodes per class (selection)

Class	1	2	3	4	5	Sum
Frequency	59	19	25	24	28	155

Total size of network (isolates included): 182
Number of non-isolated knots with class assignment: 155
Number of links among knots with class assignment: 338

a) Expected frequencies of links

	Class 1	Class 2	Class 3	Class 4	Class 5	Sum
Class 1	23.698	12.639	10.796	14.746	27.121	65.302
Class 2	10.651	5.680	4.852	6.627	12.189	34.320
Class 3	21.036	11.219	9.583	13.089	24.074	69.417
Class 4	14.911	7.953	6.793	9.278	17.065	46.722
Class 5	19.704	10.509	8.976	12.260	22.550	51.450
Sum	66.302	42.320	31.417	46.722	80.450	338.000

Chi**2: 848.182 with df = 16
Phi**2: 2.509
normalized by nrow, ncol: Cramer`s V** 0.627

b) Relative frequencies (% of expected frequencies)

	Class 1	Class 2	Class 3	Class 4	Class 5	Out (Between Classes)
Class 1	312.26	47.47	0.00	27.13	18.44	22.97
Class 2	9.39	651.35	0.00	0.00	16.41	8.74
Class 3	52.29	26.74	427.85	7.64	95.54	54.74
Class 4	13.41	12.57	0.00	549.68	11.72	10.70
Class 5	10.15	9.52	0.00	0.00	314.85	5.83
In (Between Classes)	24.13	25.99	0.00	10.70	39.78	

From a comparison of row 1 column 1 ("within class 1") with row 1 column 6 ("outgoing from class 1") we see for instance that members of class 1 named members of their own class much more frequently (312.26%) than – in the average – members of all other classes (22.97%). Among members of other classes, they named "relatively often" but less than expected by chance the members of class 2 (47.47%). A completely equal distribution of links (after controlling for class specific indegrees and outdegrees) would have lead to relative frequencies of 100% in each cell.

The data analysis in chapter 4 will follow these lines of procedure, i.e. excluding isolated students and standardizing by means of expected frequencies based on class sizes, in- and outdegrees.

3 Criteria for the description of the information flow amongst freshmen

The already mentioned data collection about talks among students about study related books and the experiment on the discounted purchase of a book offers the rare possibility to test for some consequences of a social grouping on the actual course of a communication process. Such testing is important since the identified social grouping only partly reflects the reality of the examined freshmen and results also as a consequence from our introduced definition. This refers especially to the definition of "important characteristics" in chapter 2 and to the introduced determination of the characteristics of the classification procedure and the number of classes. If we want to rely on our description of the social reality and take over the responsibility for applications of the derived findings, at least a partial testing of such intermediate steps is required.

Some limitations need first to be set out: the experiment only provides data on relatively few book purchases, and it does not provide data on the point in time on which the information on the discount purchase possibility was passed on. Therefore, we cannot trace in detail the direct or indirect course of the information flow from the first informed persons to the ones that actually purchased the book. Such data need to be derived from the available data guided by certain assumptions.
In this regard the following data are disposable:

1. Statements of the freshmen on book related talks with other students within an ongoing week:
   - "Did you buy a study related book in the last week?
     (please write here author's name and short title)
     Did you speak in this regard to any persons before the purchase?" (question no. 16)
   - "Are there other study related books about which you talked with fellow students last week?
     (please write here author's name and short title)
     Please indicate the persons you have talked with about the respective book." (question no. 17)
2. Name lists of freshmen (of course anonymous: numbers) which received by chance at certain points in time the information on the opportunity for the discount book purchase and/or bought the book at a given moment at the project group (table 4).

By looking at the length of the lists at a point in time (columns), table 4 gives at first sight an optical impression on the course of the diffusion process: as known from other studies, the number of buyers increases from week to week, because (as the explanation based on the logistic function implies) an increasing number of persons know the message and contribute to its diffusion. At a certain point the number of buyers drops - in the same logic - due to the fact that the message more and more often reaches persons that are already informed, hence, no further diffusion is affected (see e.g. A. Rapoport, 1980, chapter 3).

Having a closer look at the results it becomes evident which information is *missing*. The students 15, 78, 101, and 164 (see table 3, column 3) that bought the book in week no. 3 *do not belong to the group that was first informed*. These actors have used information they received anyhow from the students named in the first column as "first informed". But how did the information reach them?

Table 4: List of persons that received first information on the opportunity for a discount book purchase and/or actually bought the book

Starting population (Info received)		Target population book purchased in week x						
Week 3	Week 5	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9
7	6	15	43	11	10	2	36	8
26	57	78	49	13	25	24	58	64
31	62	101	76	16	38	28	115	72
43	98	164	94	26	47	35	121	125
44	100		124	29	52	97	139	
66	103		176	57	55	112	151	
87	124			84	73	123		
105	129			114	100	173		
109	130			135	116	174		
137	143			140	117			
145	144			144	128			
157	165			162	136			
163	178			170	143			
	179			181	145			
	181				153			
					154			
					156			
					174			
					175			
					178			

A first explanation is based on the assumption that the reported book talks are potential connections that might be used for the transfer of the information on the discounted book from the first informed persons to the buyers. When student A reports on a book talk with student B one may assume that information flows in both directions. This should be true even if person B doesn't report on the talk so that it has only been mentioned by one side. Results from recent examinations seem to justify doubts on this point. M. Trappmann (2004) has used the same data set to examine whether the centrality of persons in a network is related to their being informed while measuring the latter via the book purchases: the idea was that

central persons should be better informed than others and should therefore be more often amongst the buyers.

Because there are different types of centrality measures Trappmann equally stated the question which centrality concept would be most appropriate for the given problem. The result was somewhat intriguing: a centrality measure based on the outdegree (e.g. the number of book talks reported by a person) showed to be less adequate opposed to a centrality measure based on the indegree (e.g. the number of persons that have reported on a book talk with the target person). Hence, the direction of the information flow has to be assumed against the reported direction on book talks. Based on the findings of Trappmann the report or non-report about a talk may be understood as a result of the importance ascribed by the reporting person to the talk. If a person hears important news in a conversation, she will probably remember and report the situation more probably than if she has herself – maybe incidental – spoken about an issue being *important for the other person*. For the data at hand on book talks this means: if person A reports on a book talk with person B we should above all assume an information flow B→A. This is of course only of relevance for book talks that were reported from just one actor.

Apart from the direction of the information flow indicated by book talks the content of the talks is equally important for the course of the information process. Accordingly, the above stated view that all book talks are contacts for the dissemination of the information leads to an overestimation of the network size. On the other hand we know from systematic examinations on the response accuracy that a large part of contacts will not be reported at all (e.g. H. R. Bernard, P. D. Killworth, D. Kronenfeld and L. Sailer, 1985).

In a second approach we restrict the social net to contacts based on book talks where the potential "sender" of the information is in fact an "informed person" and the "receiver" a buyer (as a person that needs to be informed first). In this manner a part of the mentioned problem is solved since contacts without a close reference to the topic of the book purchase are excluded. However, still unsolved is the second problem of the incomplete knowledge of contacts relevant to the topic of the book acquisition.

Finally, due to fact that remembering earlier contacts as well as the purchase decision needs some time, the hitherto determined networks will be condensed over time: the networks from week 3 refer only to the data from week 3, the network from week 4 is assumed to consist of data from the weeks 3 and 4 (logical "or" connection), and so on. The resulting nets describe which relations in a certain week or earlier were existent between informed persons and buyers in the

sense that statements about book talks between the respective freshmen were reported.

While there are in the cumulated network of the book talks 348 (direct or 1-step) contacts, the reduction on contacts between earlier informed persons and later buyers shrinks their number to only 25 such contacts. Hence, with regard to the actual buyers – which were necessarily informed earlier about the discount purchase opportunity (!) – there is data for less than the half of these cases on the possible sources of the information. This, again, underlines the already mentioned statements on the incompleteness of reports (a.o. H.R. Bernard, P.D. Killworth, D. Kronenfeld and L. Sailer, 1985).

4 Information flow within and between social groups of freshmen

In this chapter the focus is on the significance of social cohesive groups amongst freshmen for the actual course of communication processes. On the side of the independent variable,i.e. explaining social relations (cohesive groups), we use data on expressive relations of freshmen (chapter 2). For the description of the communication process we want to explain, we use data about instrumental contacts (book talks) and the time dependent diffusion of information on the possibility for a discounted book purchase (chapter 3).

From table 4 we see that in weeks 3 and 5 a total of 13+15=28 students were informed about the discount purchase option and that 63 persons bought the book in one of the weeks (3-9) from the project group. Ten persons belonging to the group of the first informed, bought the book in one of the weeks 3-9, although one did so (in week 4) before he received the information through the project group. Taking into account the conditions on the timing only 9 buyers may, strictly speaking, be regarded as being “first informed”. Furthermore, amongst the 63 buyers one person appears in two consecutive weeks (5 and 6) so that also here in fact are only 62 buyers documented. Hence, the 53 buyers not belonging to the group of the “first informed” must have received the information either directly from the first informed or indirectly by passing through any third parties.

Before we turn to the classification of the students based on expressive contacts (see chapter 2) we will look at the relative frequencies of reported book talks (instrumental contacts) within and between the four groups which arise from the combination of the two above mentioned events defining the experimentally induced communication process:

1. neither informed nor buyers,
2. first-informed non-buyers,
3. not first-informed buyers,
4. first-informed buyers.

Table 5 describes contacts within and between these four groups (classes). The contacts are defined here by the cumulated book talks which were reported by students in the context of talks about any book - not just the discounted book! - in any of the weeks (not just in the weeks prior to the book purchase). Disregarding some differences as regards content and timing references, table 5 shows the absolute and relative frequencies of these contacts. Table 5 takes indirect relations with a maximum of 3 steps into account.

Table 5: Indirect book talk contacts within and between groups of first-informed and/or buyers (absolute and relative frequencies)

Frequencies of knots per class (total)

Class	1	2	3	4	Sum
Frequency	102	18	52	10	182

Frequencies of knots per class (selection)

Class	1	2	3	4	Sum
Frequency	90	16	49	10	165

Indirect Connections (max. 3-step)
Total size of network (isolates included): 182
Number of non-isolated knots with class assignment: 165
Number of links among knots with class assignment: 1134

a) Links between classes

	Class 1	Class 2	Class 3	Class 4	Sum
Class 1	342	41	184	18	585
Class 2	43	5	12	2	62
Class 3	205	21	180	13	419
Class 4	39	6	22	1	68
Sum	629	73	398	34	1134

b) Relative frequencies

	Class 1	Class 2	Class 3	Class 4	Out (Between Classes)
Class 1	105.40	108.87	89.62	102.62	93.28
Class 2	125.04	125.28	55.15	107.59	98.26
Class 3	88.21	77.86	122.40	103.48	87.89
Class 4	103.40	137.07	92.18	49.05	101.57
In (Between Classes)	94.25	98.54	86.87	103.25	

Chi**2: 23.537 with df = 9
Phi**2: 0.021
normalized by nrow, ncol: Cramer`s V** 0.007

Since we are – in line with our previously formulated hypothesis (see chapter 3) – interested in the information stemming from incoming relations, we concentrate first on columns. Column 3 for instance shows the percentage of respective contacts in the group no. 3, the "not first-informed buyers". One could expect that these "not first informed buyers" should have had more contacts (on average) with the informed actors – that means either with "first informed" students from classes 2 and 4 or with other buyers who had to be informed at least secondarily (i.e. with persons from their own class, or again from class 4). However, the data do not show much evidence in the direction of these expectations: there are only slightly more contacts with students from classes 2, 3, and 4 (who should have been informed directly or indirectly) as opposed to students from class 1 who were not informed directly but who possibly received the information indirectly and later. The relative frequent contacts within class 3 (122.40 %), too, do not deliver strong support for such expectations. On the other hand one could argue that the within class contacts of class 3 are much more frequent than the within class contacts of class 4, the "first informed buyers" (49.05%): The latter were informed individually and did not need any communication at all to take note about the discounted book.

So far we have analyzed the association between the first information of the opportunity for a discounted book purchase and the corresponding book purchases on one hand and the contacts based on book talks on the other hand. We did not find any convincing evidence for strong relationships. However, the main focus here is on the assumption that the process of information diffusion in the freshmen population wasn't uniform: it should differ with regard to the kind of expressive relations realized in the different groups. The explanation of such transmission processes in the context of easily identifiable groupings makes also sense from an

application perspective, since complex process data surveys like the ones used here on freshmen and the course of information diffusion over 6 weeks are rather an exception. For practical purposes such a setting would most probably proof to be prohibitive.

Yet, such groupings don't need to be based on expressive relations but may refer to common opportunity structures in the context of institutional contacts (e.g. on the number of commonly visited courses at university) or other significant characteristics of actors. Though the question of which such characteristics may be seen as especially suitable for the identification of groupings that are important for a process can't be answered by our findings since we restrict ourselves to the analysis of the appropriateness *of a single* grouping.

Table 6: Density of contacts (book talks) within and between 5 groups (based on expressive relations)

Frequencies of knots per class (total)

Class	1	2	3	4	5	Sum
Frequency	86	19	25	24	28	182

Frequencies of knots per class (selection)

Class	1	2	3	4	5	Sum
Frequency	71	19	24	23	28	165

Indirect Connections (3-step)
Total size of network (isolates included): 182
Number of non-isolated knots with class assignment: 165
Number of links among knots with class assignment: 1134

a) Links between classes

	Class 1	Class 2	Class 3	Class 4	Class 5	Sum
Class 1	158	37	20	47	58	320
Class 2	6	159	1	0	1	167
Class 3	38	5	58	12	91	204
Class 4	33	0	53	97	34	217
Class 5	29	15	37	5	140	226
Sum	264	216	169	161	324	1134

b) Relative frequencies

	Class 1	Class 2	Class 3	Class 4	Class 5	Out (Between Classes)
Class 1	212.09	60.70	41.94	103.45	63.44	65.99
Class 2	15.43	499.85	4.02	0.00	2.10	5.92
Class 3	80.01	12.87	190.78	41.43	156.13	84.10
Class 4	65.32	0.00	163.89	314.85	54.84	64.45
Class 5	55.12	34.85	109.85	15.58	216.81	53.27
In (Between Classes)	55.94	30.95	80.09	49.16	70.93	

Chi**2: 1200.497 with df = 16
Phi**2: 1.059
normalized by nrow, ncol: Cramer`s V** 0.265

Table 6 informs about the importance of the chosen grouping of expressive relations (chapter 2) for the connections based on book talks. Again, the latter have been combined over all survey weeks.

In some but not all cases the contact densities based on the book talks (links with up to 3 steps) reflect a similar image like the one we obtained in chapter 2 while analyzing the 5 groups with regard to the existing one step expressive ties (see also table 3): all relative frequencies of instrumental within class contacts in table 6 in the diagonal exceed clearly the expected values (which are 100%, see chapter 2). This holds especially – as with expressive contacts – classes 2 and 4 (499.85% and 314.85% of the expected values). Complementarily, and partly as a consequence of the form of standardization used here, students of class 2 do not have many instrumental contacts with students of other classes. The contacts of students in class 2 are almost exclusively restricted on their own class.

On the other side, students of class 4 report more outside contacts to class 3 (163.89%) than expected while receiving approximately the expected frequency of contacts from class 1. Both characteristics differ from what we noticed in the context of expressive contacts (table 3). The relative frequencies of instrumental within-class contacts of classes 1, 3, and 5 are also much higher than the expected values, even though smaller than those of classes 2 and 4. Worth mentioning are furthermore the contacts of students from class 3 with class 5 (especially outgoing: 156.13%; to a lesser degree incoming: 109.85%), and with class 4 (incoming: 163.89% but not outgoing: 41.43%). With expressive contacts a similar relation may be found from class 3 (outgoing) to class 5.

As an intermediate result we conclude that the groupings obtained from analyzing solely the expressive relations clearly reflect the structure of instrumental contacts (book talks). This, in turn means, that data on expressive relations which are relatively easy to collect provides equally good information about instrumental contacts.

In the following last step the perspective will be restricted to the target criteria already mentioned in chapter 3. The focus is now on the book talks while we are also taking into account the timing conditions of these directed relations, limiting the analysis to such contacts where the assumed "sender" was in fact informed about the purchase opportunity at the point in time of reporting on this possibility and the "receiver" of the information actually bought the book later on. Because under this perspective only but a few ties are left for the analysis we again included indirect ties up to a length of 3 steps that were combined over all weeks.

Table 7: Connections between first-informed students and buyers based on book talks within and between (expressive) groups (combined over weeks 3-9)

Frequencies of knots per class (total)

Class	1	2	3	4	5	Sum
Frequency	86	19	25	24	28	182

Frequencies of knots per class (selection)

Class	1	2	3	4	5	Sum
Frequency	26	4	17	5	8	60

Indirect Connections (3-step)
Total size of network (isolates included): 182
Number of non-isolated knots with class assignment: 60
Number of links among knots with class assignment: 201

a) Links between classes

	Class 1	Class 2	Class 3	Class 4	Class 5	Sum
Class 1	22	3	17	7	15	64
Class 2	0	2	0	0	0	2
Class 3	17	5	29	6	28	85
Class 4	2	0	7	6	6	21
Class 5	6	2	8	2	11	29
Sum	47	12	61	21	60	201

b) Relative frequencies

	Class 1	Class 2	Class 3	Class 4	Class 5	Out (Between Classes)
Class 1	147.01	78.52	87.53	104.69	78.52	85.65
Class 2	0.00	1675.00	0.00	0.00	0.00	0.00
Class 3	85.53	98.53	112.42	67.56	110.35	94.59
Class 4	40.73	0.00	109.84	273.47	95.71	79.76
Class 5	88.48	115.52	90.90	66.01	127.07	88.48
In (Between Classes)	78.04	84.17	90.90	79.76	95.44	

Chi**2: 49.047 with df = 16
Phi**2: 0.244
normalized by nrow, ncol: Cramer`s V** 0.061

Apart from the overall weak relational density – which is the result of the strict definition of the criteria variable – we find a pattern for relative frequencies in table 7 which is very similar to the one in table 6. This relative stability and the few extreme values that are presumably depending on random errors due to the very small frequencies support our conclusions and assure that the grouping on the basis of expressive relations allows equally for insights into the process of an instrumentally based information flow – at least with regard to the case at hand.

References

Bernard, H. R., P. D. Killworth, D. Kronenfeld & L. Sailer. (1985). On the validity of retrospective data: The problem of informant accuracy. In: *Annual Review of Anthropology,* 13, 495-517.

Hummell, H. J. & W. Sodeur. (1984). Interpersonelle Beziehungen und Netzstruktur. Bericht über ein Projekt zur Analyse der Strukturentwicklung unter Studienanfängern. In: *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 36(3), 511-556.

Projektgruppe Studienanfänger. (1982). Strukturentwicklung und Informationsprozesse in einer Population von Studienanfängern. 1.1 Datenerhebungsbericht; 1.2 Anlagen; Hans Joachim Hummell und Lothar Krempel, Universität Duisburg, Klaus Echterhagen und Wolfgang Sodeur, Universität Wuppertal, (financed by the German National Science Foundation (DFG), Nr. HU 288/2.

Rapoport, A. (1980). *Mathematische Methoden in den Sozialwissenschaften*. Würzburg/ Wien: Physica.

Sodeur, W. (1974). *Empirische Verfahren zur Klassifikation*. Stuttgart: Teubner.

Sodeur, W. & V. G. Täube. (2005). *Formation of Social Groups and Information Flow amongst Freshmen* (Original in Russian: Перспективы: Сборник научных статей аспирантов; Выпуск 4), Niznij Novgorod.

Späth, H. (1975). *Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion.* München/Wien: Oldenbourg); engl. 1980, New York.

Täube, V. G. (2003). Social Capital and the Evolution of Social Structures. *Presentation at the International Social Network Conference (SUNBELT XXIII.)*, Cancun, Mexico, 12.-16. February, 2003.

Täube, V. G. (2004). Measuring the Social Capital of Brokerage Roles. In: *Connections*, 26(1), 29-52.

Trappmann, M. (2004). Centrality in Student Networks and the Reception of Relevant Information. *Presentation at the International Social Network Conference (SUNBELT XXIV.)*, Portoroz, Slovenia, 12.-16. May, 2004.

Wasserman, S. & K. Faust. (1994). *Social Network Analysis. Methods and Applications*. Cambridge: Cambridge University Press.

Academic Employment Networks and Departmental Prestige

Debra Hevenstone

Abstract Research has found a correlation between academic departments' rank and their centrality in academic hiring networks. This correlation results from the fact that highly ranked schools train more PhDs, their graduates are more likely to find first jobs in academia, and that they have more faculty. This study is the first to consider this correlation independent of training and department size. One expects no correlation because mid-career academics move between institutions for a variety of reasons such as wages, location, and specialty areas. Nevertheless, this study finds that the correlation persists; suggesting individuals are more willing to make career switches to top departments or between them. This gives top departments a competitive advantage and positive returns to their rank, with their faculty disproportionately linked to institutions and researchers at other departments. This could be one reason for the stagnancy of academic rankings.

Acknowledgments Many thanks to Lada Adamic, Yu Xie, Mark Mizruchi, Stefan Wherli, and Jonas Nart as well as an IGERT fellowship from the University of Michigan Center for the Study of Complex Systems . All errors are, of course, mine.

1 Introduction

Academic rankings incorporate both "objective" measures of department quality (such as citation rates and funding patterns) as well as subjective measures. While we might expect that objective measures would allow departments to improve their rankings, academic rankings are relatively constant over time, with the top schools swapping the top positions (Graham and Diamond 1997). Several rankings for sociology graduate programs from 1925 to 2005 are illustrated in table 1.

Four institutions have been in the top 10 since 1925 while 8 others have since 1982. Despite their consistent results, these rankings use significantly different methodologies to come to their results. The oldest type of formula is a reputational rank, which was pioneered by Raymond M. Hughes. In his 1925 report Hughes surveyed 20 to 60 faculty members in each field, asking them to rank institutions based on "esteem at present time for graduate work in your subject." The much critiqued US News and World Report rankings build on this formula, basing ranks on a peer assessment surveys (50% response rate) sent to academic department heads and directors of graduate study in sociology.[1] The National Research Council's (NRC) 1995 rankings are more complicated; they also use reputational measures (also with about a 50% response rate) but augments it with data for about 17 program characteristics such as: size, private vs. public, total research and development (R & D), federal R & D, library expenditures, enrollment, total faculty, % faculty FT, % faculty with research support, percent full professors, faculty awards, awarded faculty, citations per faculty, faculty characteristics, and student characteristics. The NRC found that the reputational measures were consistent with the objective measures. Critiques of the NRC rankings argued that there was too much emphasis on research-related variables and too little on doctoral training.

Table 1: Sociology Department Ranks

1925*	1982+	1995**	1995+	2005**
Chicago	Chicago	Chicago	Chicago	Wisconsin
Columbia	Wisconsin	Wisconsin(2/3)	Wisconsin(2/3)	Berkeley
Wisconsin	Berkeley	Berkeley(2/3)	Berkeley	Michigan(3/4)
Minnesota	Michigan	Michigan(4/5)	Michigan	Chicago(3/4)
Michigan	Harvard	Chapel Hill(4/5)	UCLA	Chapel Hill
Harvard	Chapel Hill	Harvard(6/7)	Chapel Hill	Princeton(6/7)
Missouri	Stanford	UCLA(6/7)	Harvard	Stanford(6/7)
-	Columbia	Stanford	Stanford	Harvard(8/9)
-	UCLA	Northwestern(9/10)	Northwestern	UCLA(8/9)
-	Arizona	Princeton (9/10)	Washington	UPenn

*Hughes (1925)
+National Research Council (1982, 1995)
** US News and World Report (1995,2005)

[1] For an excellent critique of the US News rankings see Ehrenberg 2002

This analysis relies primarily on the NRC's sociology rankings, when including foreign institutions in the analysis; I also use the Newsweek international rankings (not specific to sociology).[2] The Newsweek score includes measures of citations, publications, international faculty, international students, faculty:student ratios, and library holdings. While the two rankings are developed using different metrics, the rankings correlate at .625 for those US schools where both ranks were available. The primary difference between the ranks is that the NRC sociology rankings exclude technical/science schools like MIT and Caltech, while these schools are near the top of the general international ranking.

Some researchers suggest the stagnant rankings indicate a closed system where departments find it difficult to move up the rankings and where well-established programs can reinforce their dominance. This organizational situation could be considered analogous to individual-level stratification in a "closed system" where intergenerational transmission of advantage trumps equal opportunity (Lipset et al. 1955). Ideally stratification should function as an incentive for individuals to work harder or acquire more human capital (Davis and Moore 1945) or for organizations to innovate and improve their product. However, too much stratification might indicate that either individuals are able to earn more based on their current assets or analogously, an organization can sell more of their product not based on their current effort but on their inheritance.

There are two reasons that the rankings might remain stagnant. First, it might be that respondents to the reputational survey are rather ill informed, basing their evaluation of doctoral programs not based on their actual merit but on what they have heard. If this is the case, once a program is highly ranked, it will remain there, as professors perpetuate the reputation without objectively examining it. Second, once a program is highly ranked, it has resources to perpetuate that rank by attracting faculty. This second reason is the basis for this paper. While some lament the caste system, the simple preference for faculty to move to or between higher ranked schools can cement departments' central position. This central position can translate into departmental prestige through many mechanisms not explored in this paper- such as research collaborations, knowing about upcoming trends in the field, or hosting small conferences leading to publications. While this paper does not explore the mechanisms linking hiring network centrality and prestige, it does confirm the existence of the correlation independent of training and department size. While departments' central positions might aid their faculty, the pattern of the highest ranked department sitting at the centre of a hiring network

[2] I tested the Shanghai rankings as well, though there was little difference.

probably gives the department an advantage as well and is a likely an example of positive feedback.

There is significant non-network research testing how the institutional prestige of PhD granting institutions influences first job placement. The literature finds that the most prestigious universities hire each other's graduates, over-valuing the institutional prestige of applicants' training institution over their other characteristics that might be more predictive of success, such as the time it took to complete the PhD (Bair 2003, Baldi 1995, Burris 2004, Burke 1988, Hargens and Hagstrom 1966, McGinnis and Long 1988, Reskin 1979, Smelser and Content 1980).

In contrast, there are only four papers testing whether academic departments' positions in academic hiring networks is linked to academic rank. Burris (2004); Wiggins et al. (2006) and Fowler et al. (2007) tie professors to their current employers and their PhD granting institutions, generating a network of institutions with weighted, directed ties indicating the number of PhDs trained at one department and currently employed in another. These studies analyze computer science, information, sociology, and political science departments and find a significant relationship between network centrality and rank for all of them. Their choice of centrality measures vary, though they all use recursive network measures (based on the adjacency matrix's dominant eigenvector) that measure a node's prestige based on the prestige of those nodes it is connected to. Centrality measures used include: eigenvector centrality (Bonacich 1972), PageRank (Page et al. 1999), and hub and authority centrality scores (Kleinberg 1998) (used by Burris 2004, Wiggins 2006, and Fowler 2007, respectively). Fowler et al. (2007) uses hubs and authorities, making a distinction between prestige from placing students at prestigious departments and hiring professors from prestigious departments. All three studies ignore the link between the department where an academic got their PhD and the department of their first job (the traditional question in the non-network studies) and ignore all placements between current job and training. Grannis' (2005) approach is slightly different, looking at UCLA's ego network of its faculty trades with other departments. These articles then use node centrality as a predictor of departmental prestige (Burris 2004, Fowler et al. 2007) and often interpret the relationship as confirming institutional stratification (Burris, 2004) or as showing that placing students in prestigious schools is more relevant to prestige than hiring professors from prestigious schools (Fowler et al. 2007).[3] Ultimately, it

[3] Using PageRank as a predictor of academic Rank Wiggins finds coefficients as high as 11.2. Burris finds coefficients around 1.3 in Sociology, History, and political science, for a ln transform of eigenvector centrality. Finally, Fowler et al report correlations as high as .82 between predicted rank based on PhD exchange networks and actual rank. All three have sig. findings.

is difficult to parse the relationship out since there is a circle of causality- productive researchers increase a school's prestige, but prestigious schools also attract researchers.

This paper expands on the current body of research in two ways. First, it considers the impact of training many more PhDs than there are openings for professors (henceforth referred to as "overtraining"), and second, it considers the spurious effect of department size – a reliable predictor of both hiring network centrality and academic rank.

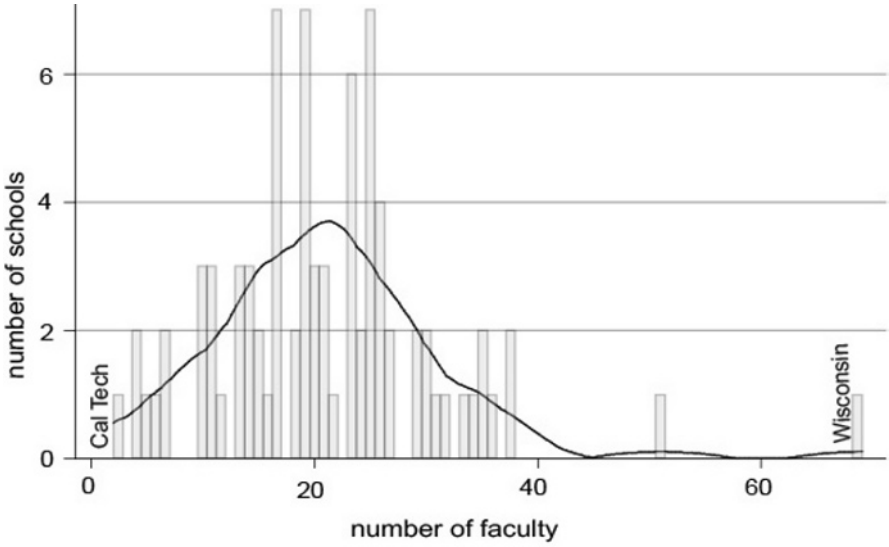


Figure 1: Department Size

Currently, the literature ignores that centrality and prestige are both strongly influenced by department size. As described earlier, the NRC rankings are based on both reputation and objective measures. One of these measures is department size (the number of faculty and students) (National Research Council 1982, 1995). Department size also indirectly influences the rankings insofar as there are more former employees and students from the largest schools, assuming that individuals rank prior affiliations higher. Department size also increases centrality because bigger departments have more ties. Consequently, centrality and prestige should be correlated by virtue of department size even if location in the network is unrelated to prestige. This is well illustrated in one of the four existing studies, Fowler

et al. (2007), who shows that ranks can change when we consider department size, particularly for boutique programs with focused research areas. Theoretically we might consider department size to play a valid, not a spurious, role since bigger departments have more depth and thus more opportunities for graduate students and researchers to expand their skills. As such, size is an indicator of program quality.

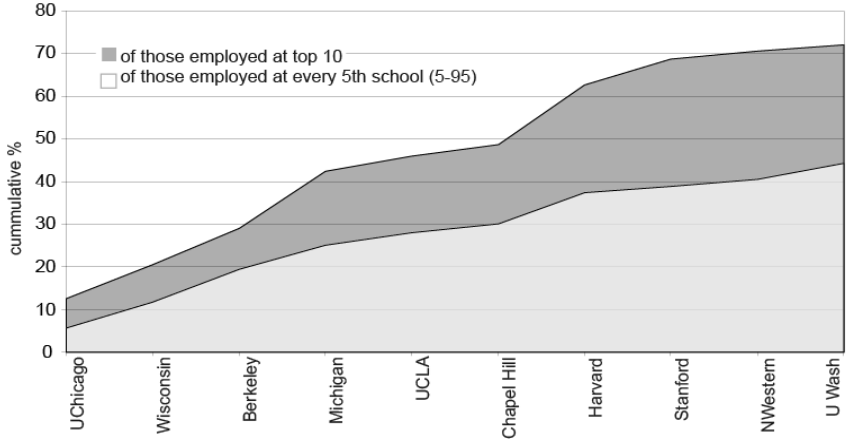


Figure 2: Where professors were trained

Overtraining can also account for part of the relationship between centrality scores and academic rankings. The current research ties professors to their training departments and to their current department. If the most prestigious and largest departments train a much larger percentage of the job market than they hire, and than the less prestigious schools, they will be more central. Figure 2 shows the proportion of professors trained at the top 10 schools. The grey section shows the proportion of professors currently employed at the top ten schools who were also trained at the top ten schools. We see at the origin, that over 10 percent of faculty at top ten schools were trained by the University of Chicago and about 20% were trained at Wisconsin and Chicago combined. Bumps in the graph show that University of Michigan and particularly Harvard graduates are more likely to be at top schools. Over 70% of the professors at top ten schools were also trained in the top 10 schools. The lower line and the light grey section of the graph show where the professors at every fifth school in the rankings were trained, from the 5th to the 95th school. This line shows the same pattern as that for the top ten schools, with close to half of all professors being trained at the top ten schools. The ASA reports

there are 598 new PhDs every year and only 4,227 tenure and tenure track positions in the US; the entire profession could be replaced every 7 years, meaning that all universities can hire from the top schools, placing highly ranked schools at the centre of the hiring network. This analysis takes training into account, testing whether the association between hiring network centrality and rank holds independent of training.

2 Data and Methods

Two separate data sets were collected; each was collected by selecting sociology departments, going to their web sites, collecting the CV's of current permanent faculty, and entering ties between the faculty and organizations they had been affiliated with in the past. The first data set collected faculty from prestigious departments (Wisconsin, University of Michigan, Harvard, Berkeley, UCLA, University of Chicago, Brown, Stanford, and University of Arizona). The second sample was collected with the intention to validate the effect of having sampled the most prestigious institutions in the first data set. This second group includes: Yale, University of Pennsylvania, Northwestern, Princeton, Johns Hopkins, and NYU. The second group was chosen to represent a still exceptional, although not top, schools with the intention of testing whether these schools became the most important when they were sampled. Surprisingly, they did not. One tie was coded for each of the faculty's current and past institutional affiliations, modeling full career paths. Ties were then coded as PhD training institution, tenure-track jobs, and non-tenure track jobs. "Non-tenure track" jobs include lecturers, post-doctoral, non-academic, and visiting appointments. Approximately 7% of the sample did not have their CV's posted on-line. For these cases, ties were coded to the faculty's current institution and their PhD granting institution (which was normally listed). The samples included 193 and 241 institutions, 99 and 89 institutions that were ranked by the NRC, and a total of 886 and 882 ties for samples one and two respectively. All network measures used in this analysis were generated using the full graphs, although the secondary regression analyses use the sub-sample with academic rankings.

I analyze 12 different graphs, testing whether the relationship between centrality and prestige is robust to graph specification. The graphs analyzed differ along three dimensions: sample choice (2), ties included (3), and graph reduction (2). The graphs either: included all three types of ties, excluded non-tenure track ties, or excluded training ties. The first sample was reduced to 99 institutions when

non-tenure track ties were excluded and to 178 institutions when student ties were excluded while the second sample was reduced to 89 and 237 institutions. Each of the 6 graphs was then reduced to include only institutions, weighting the ties between them by the number of people had in common.

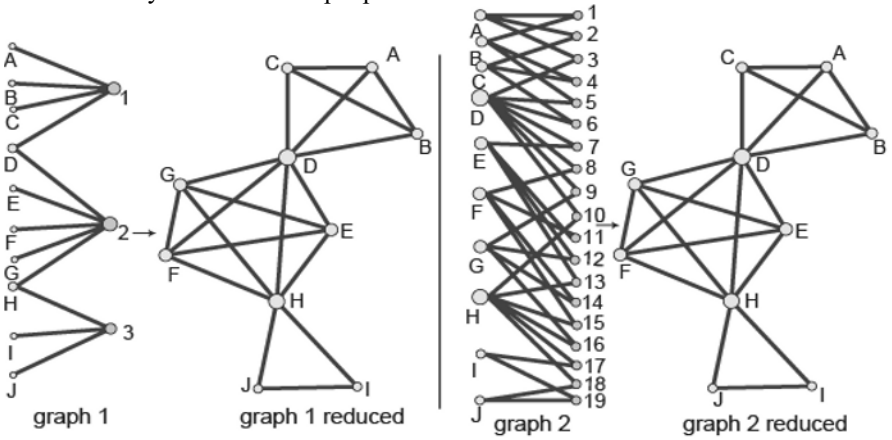


Figure 3: Reducing two different bipartite graphs into one reduced graph

There are 4 main methodological challenges using this data. First, any sampling method biases the graph, enhancing the sampled institutions' centrality. One solution to this problem is to start with the seed institutions, and then sample from the other institutions that enter the analysis (snowball), ultimately excluding the original seed institutions from the network analysis. Instead, I include these biased observations, but use two different seeds, concluding that if the results are similar using the two seeds, the conclusions are robust to sample bias. Second, the data includes both end-of-career and beginning of career professors. This biases the data insofar as older professors with a longer history of institutional connections are more likely to be at more prestigious universities. Other studies have similar problems, for example, coding the tie between a department that trained a professor and their first job the same as their emeritus job (Burris, 2004; Wiggins et al., 2006; Fowler et al., 2007). Third, academia is not an isolated network, which can bias network statistics like transitivity, degree distribution, and clustering (Grannis, 2005) as well as mean degree (Kossinets, 2006). The final difficulty is that the graph is bipartite with two types of nodes (professors and departments) linked by ties (employment relations). Bipartite graphs are also called an "affiliation networks." Most centrality measures are designed for one-mode graphs (Borgatti and Everett, 1997) but can easily be adjusted for use with bipartite graphs, or the orig-

inal centrality measures can be used on the reduced form of the bipartite graph. Centrality measures (defined in the following section) differ substantially based on the approach taken. Figure 3 shows two graphs that are different in their bipartite forms but identical in their reduced forms. Graph 1 could be a picture of three professors who have had very mobile careers, while graph 2 illustrates 19 professors, each of whom is only affiliated with their training institution and their current employer. In the reduced versions of the graphs D and H are the most important nodes, while they are more important in bipartite graph one than two. Calculating the nodes' centralities, D and H have similar eigenvector centralities in both graphs. However, D & H have much higher standardized degrees and closeness centralities in both graph 1 and the reduced graph than in graph 2.[4] As such, I analyze the graph both as a bipartite and a reduced graph using the bipartite centrality measures proposed by Borgatti and Everett (1997) and illustrated in Robins and Alexander (2003) (although eigenvector centrality need not be adjusted for the bipartite graph (Bonacich, 1972, Faust, 1997)).[5]

Three different centrality measures were calculated: closeness, degree, and eigenvector centrality. Eigenvector centrality was chosen as the recursive measure, closeness centrality chosen as a distance measure (related to how quickly the department can access information from peers about funding, new research trends, recruiting, etc), and degree centrality was chosen as a straw man (it should capture department size and the experience of the department's faculty) as it should be the most biased for the sample seed. Surprisingly, results are similar using all three measures.

Standardized degree, equation 1, measures the percent of possible connections that an institution has (to institutions in the reduced graph or to professors in the bipartite graph). In both cases, the numerator is the raw degree, and the denominator is the maximum number of possible connections in the graph, np = (the number of professors) in the bipartite graph and nd-1 = (the number of departments less the department being considered) for the reduced graph.)[6] Degree centrality measures a combination of department size (faculty and training depending on the graph) and the department's turnover rate.

---

[4] Centrality scores for D in bipartite graph 1 are: .377(eig), .667(degree), .889(closeness); in bipartite graph 2 they are: .469(eig), .368(degree), .836(closeness); in the reduced: .490(eig), .778(degree), .818(close)

[5] Centrality measures for affiliation networks are also covered in Faust, 1997. I use several of the methods detailed in Faust including equation 18 to calculate eigenvector centrality. Closeness centrality is taken from Borgatti.

[6] Standardized degree was calculated using the unweighted version of the reduced graph.

$$D_i^{sb} = \frac{D_{rb}}{n_p} \quad , \quad D_i^{su} = \frac{D_{ru}}{n_d - 1} \tag{1}$$

Closeness centrality measures the inverse of the average distance between a given node and all other nodes and is illustrated in equation 2. Here, i is the node of interest, n_j indicates the number of nodes, and D_{ij} is the distance from node i to node j. The bipartite graph measure multiplies the average inverse distance by 2 to account for the fact that all connections between institutions are twice as far as in the reduced graph.[7] Closeness centrality measures whether actors can contact on another through short paths (Faust, 1997).

$$C_i^b = 2 * \frac{n_j - 1}{\sum_{j=n}^{j=1} D_{ij}} \quad , \quad C_i^u = \frac{n_j - 1}{\sum_{j=n}^{j=1} D_{ij}} \tag{2}$$

Eigenvector centrality Bonacich (1972), is a recursive measure of prestige related to Page Rank (Page et al. 1999), hubs and authorities (Kleinberg 1998), and SALSA (Lempel and Moran 2000). All are based on the dominant eigenvector of the graph's adjacency matrix and all gauge the importance of a node based on the importance of its neighbors. Page Rank adds a damping factor to the adjacency matrix (reducing the ties in the adjacency matrix by some small amount and then adding uniform random ties from each node to all other nodes) and then calculates eigenvector centrality. Both SALSA and hubs and authorities use the dominant eigenvectors of the adjacency matrix times it transpose (and vice versa) with SALSA using row and column standardized versions of the adjacency matrix. For eigenvector centrality, given the adjacency matrix, A where entry A_{ij} is 1 or 0 in the bipartite graph, or the number of connections between institutions in the reduced graph, the centrality score, c_i, is:

$$c_i = \alpha \sum A_{ij} c_j \tag{3}$$

where c is the eigenvector paired with A's largest Eigen value, i.e. the principle eigenvector. For the bipartite graph, eigenvector centralities for individuals are simply dropped.[8] All centrality scores were converted into ranks to be comparable

[7] Closeness centrality was also calculated using the unweighted version of the reduced graph.

[8] Page Rank and Hubs and authorities were also tested, yielding similar results.

to academic rank. There were no ties because the base centrality measures are continuous.

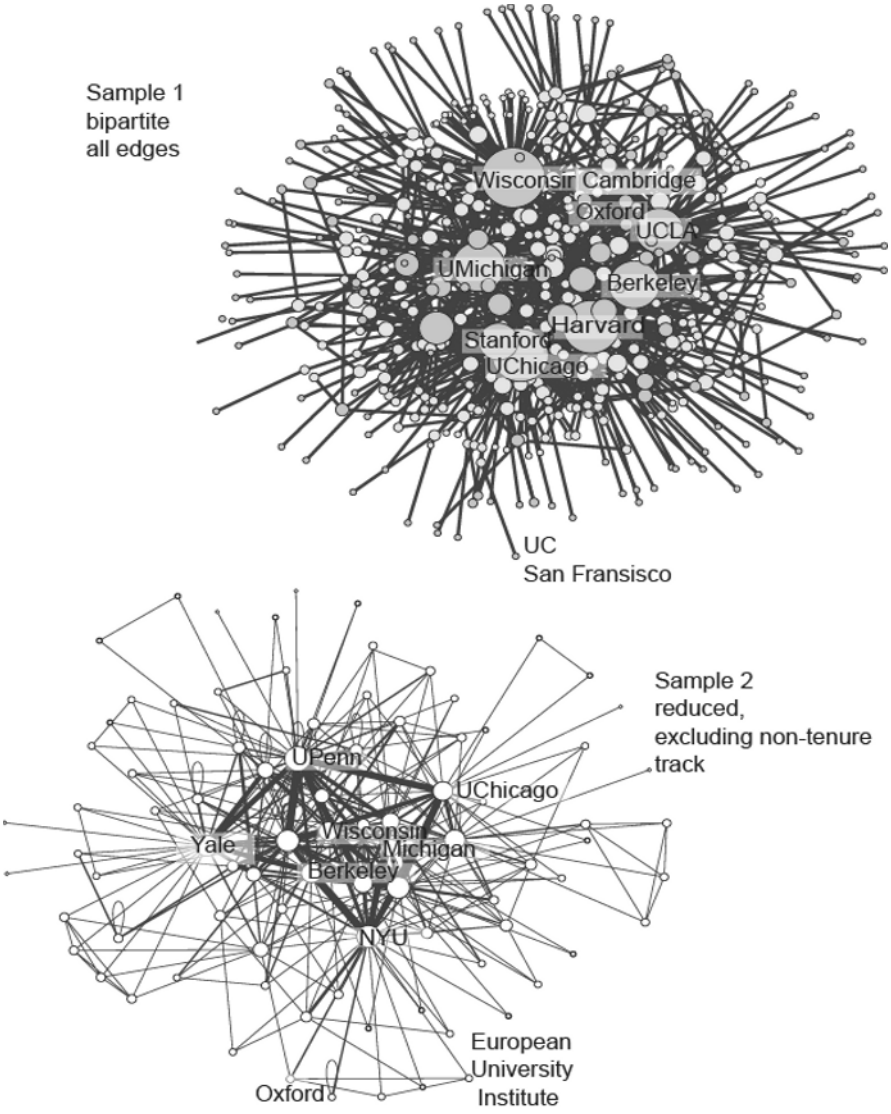


Figure 4: Two sample graphs

There are three exogenous variables: the domestic and international ranks described in the first section of the paper, and department size. For domestic universities department size was taken directly from the NRC report when possible, and from departmental web sites when not. Information was drawn from departmental web sites for non-US universities.

Two of the twelve networks are depicted in figure 4 using the Kamada-Kawai spring layout algorithm. This algorithm places "springs" between each pair of connected nodes, and moves the nodes to minimize the springs' energy. Thus, nodes are connected in clusters with the nodes they share many connections with.

I present just 2 of the 12 graphs for the sake of brevity. The first graph in figure 4 is the bipartite graph from the first sample (the very prestigious sample), including all ties (tenure, non-tenure, and student). The size of the nodes indicates their degree and the shade indicates whether they are an institution (darker) or an individual (lighter). Sampled institutions, of course, have high degrees and are central while European English-speaking institutions are also central but with smaller degrees. The halo of small institutions indicates small departments like UCSF (labeled) or non-profit and public institutions. The second graph in figure 4 is sample two's reduced graph excluding non-tenure track ties. The institutions that were part of the first sample remain central, though less dominant as they were not the sample's seed, while sampled institutions like Yale take a more dominant position. In the analyses excluding non tenure track ties foreign institutions either dropped out of the graph or moved to the periphery. Self-ties (indicating that an institution had two relationships with the same individual i.e. training and then employing the same person) become apparent in the second graph because it is sparser. Removing student ties as well, the traditional central institutions remain central though not disproportionately so.

Table 2 shows the descriptive statistics for all graphs. Average degree indicates the average number of individuals the department is associated with in the bipartite graphs and the average number of institutions sharing connections to professors in the reduced graphs. Average distance measures the average number of jumps to get from one institution to another for the reduced graphs and institution-individual-institution jumps for bipartite graphs. Finally, diameter measures the longest shortest path between any two institutions, which is 4 in all graphs. The reduced graphs have higher average degrees and lower average distances, as departments are tied to most other departments. There is no marked difference between the two samples. Comparing graphs by tie inclusion, bipartite graphs' degree increases excluding non-tenure track jobs because peripheral institutions drop

out. Removing student ties, average degree decreases because few nodes drop out but many ties do.

Table 2: Sociology Department Ranks

	All nodes	Org nodes	edges	Avg degree	Avg distance
sample 1					
bipartite all edges	479	193	886	4.59	1.92
reduced all edges	193	193	952	9.87	2.30
bipartite no non-tenure	386	99	642	6.57	1.73
reduced no non-tenure	99	99	321	6.48	2.35
bipartite no student	4.57	178	631	3.56	2.08
reduced no student	178	178	631	7.97	2.45
sample 2					
bipartite all edges	425	241	882	3.66	1.98
reduced all edges	241	241	1712	21.83	2.28
bipartite no non-tenure	273	89	509	5.79	3.83
reduced no non-tenure	89	89	331	7.44	2.37
bipartite no student	421	237	700	2.95	2.07
reduced no student	237	237	1533	12.9	2.35

Diameter for all graphs was 4

3 Analysis

Ranks generated from centralities are strongly correlated to prestige though the relationship varies across graphs. Closeness centrality changes when the graph is reduced, eigenvector centrality changes when ties are removed, and mean degree changes both when the graph is reduced and when student or non-tenure track ties are excluded.[9]

Using equation 4 to calculate the sum square deviations between the predicted and actual ranks for the top universities, we assessed which graph's centrality scores best predicted academic rank. Gs is the graph's sum of squared errors, u is a university, r is u's NRC rank, and e, c, and d are the eigenvector, closeness, and degree centrality ranks respectively,

[9] All the listed changes are significant at a 95% confidence level.

$$G_s = \sum_{u=1}^{u=10} [(e-r)^2 + (c-r)^2 + (d-r)^2] \quad (4)$$

The bipartite graph from sample one excluding non-tenure track ties was the best predictor of academic rank while the reduced graph from sample one including all ties was the worst predictor. The first three columns of table 3 show ranks generated from the three centrality scores for the best graph and the second shows those from the worst graph. In the last two columns, the average rank is in bold if the school was part of the sample seed. It seems that departments have higher ranks as predicted by centrality scores when they are part of the seed, but the top schools remain highly ranked even if left out of the seed. The graphs excluding student ties have significantly worse predictions of rank, with 3 of the 4 graphs excluding student ties landing in the bottom four (of 12) predictions.

Table 3: Centrality rankings for the best and worst graphs

	best graph			worst graph			across graphs	
	Eigen rank	close rank	degree rank	eigen rank	close rank	degree rank	Sample one	Sample two
Chicago	2	2	2	2	3	3	2	12
Wisconsin	1	1	1	2	1	1	1	8
Berkeley	5	4	5	4	6	5	5	5
Michigan	4	5	4	1	4	2	3	13
UCLA	6	6	6	5	2	6	4	14
Chapel Hill	15	15	13	55	21	16	15	15
Harvard	3	3	3	57	5	4	6	7
Stanford	7	7	7	6	7	7	7	9
Northwestern	11	10	12	10	12	12	11	4
Washington	37	28	29	81	35	45	41	23

best= sample 1, bipartite, non non-tenure edges; worst= sample 1, reduced graph, all edges

The three centrality measures seem to all predict prestige well in table 3 because they are closely correlated to one another as illustrated in table 4. The first column shows the correlation between eigenvector and closeness, the second shows eigenvector and degree and the third shows closeness and degree. The main entries indicate rank correlation while the numbers in parentheses are the correlations for the raw centrality scores. While the first 12 rows illustrate the correlations within graphs, the last line is the overall correlation, ignoring graph specification. All the

rank correlations are better than the ones using raw centralities. The graph with the most inconsistent centrality scores is the reduced graph from sample one with all ties. The ranks generated by the three centrality measures are similar regardless of graph specification.

Table 4: Correlations between centrality measures by graph type

	graph type				
sample	reduce	edges	eig-close	eig-degree	close-degree
1	yes	all	.614(.733)	.589 (.888)	.889 (.840)
1	yes	PhD & tenure	.915(.866)	.948 (.989	.922 (.878)
1	yes	no PhD	.929(.891)	.935 (.977)	.927 (.870)
1	no	all	.979(.759)	.865 (.990)	.836 (.792)
1	no	PhD & tenure	.983(.681)	.935 (.981)	.930 (.732)
1	no	no PhD	.949(.648)	.787 (.873)	.827 (.810)
2	yes	all	.959(.828)	.931 (.987)	.958 (.863)
2	yes	PhD & tenure	.977(.897)	.955 (.944)	.969 (.951)
2	yes	no PhD	.975(.854)	.940 (.924)	.955 (.928)
2	no	all	.985(.811)	.903 (.988)	.905 (.817)
2	no	PhD & tenure	.961(.721)	.936 (.974)	.952 (.764)
2	no	no PhD	.908(.767)	.841 (.944)	.924 (.696)
overall			.877(.650)	.913 (.523)	.911 (.795)

Entries are rank correlations

(…) are continuous correlations

All the centrality measures have strong correlations with domestic and foreign academic rank. For domestic ranks, the ranks generated using eigenvector centrality have a .68 rank correlation compared to .72 using closeness and .73 using degree. Correlations are slightly lower (.55, .59, and .59) for foreign international academic rank. Correlations varied substantially within the individual graphs. For example, ranks generated from eigenvector centrality had a correlation with domestic prestige ranging from .39 for sample 1's bipartite graph including all ties to .8 for sample 1's reduced graph excluding non-tenure ties. Closeness rank was somewhat more consistent with correlations ranging from .6 to .8 and from .62 to .77 for degree ranks' correlations.

In terms of biases for the seed institutions; for the first sample both eigenvector and closeness centrality under-ranked the sampled departments, though closeness centrality did less so. This is the opposite of what we expected, anticipating that

eigenvector centrality would be a more resilient estimate of academic ranking. For the second sample of less prestigious schools, using closeness centrality, seed institutions were ranked on average 10.6 positions higher than their NRC ranks and 9.28 positions too high using eigenvector centrality. (The mean rank for the sampled institutions in sample 2 was 3-7 (95% confidence) while the NRC mean rank was 13-16 (95% confidence). Because the top institutions were sampled in the first sample it was impossible to over-rank them, but in the less prestigious second sample we find the anticipated bias, with the recursive centrality measure no more resilient than closeness centrality.

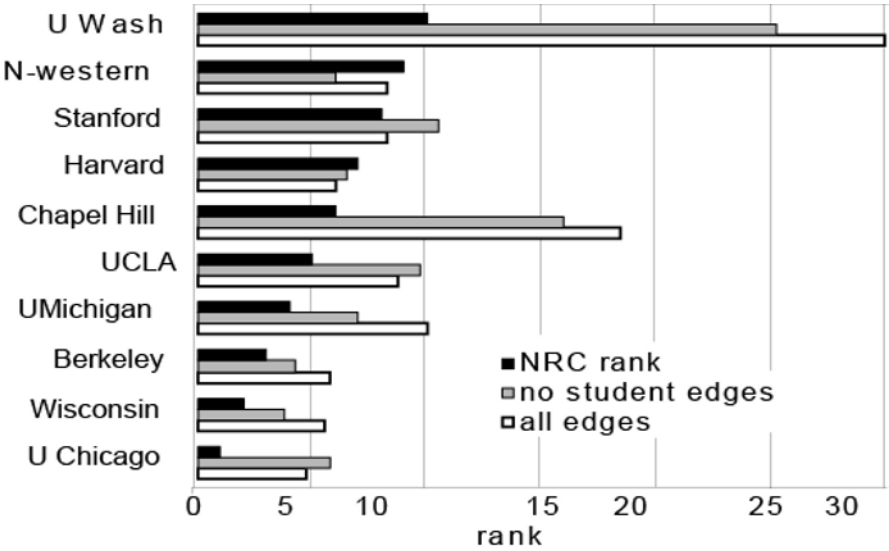


Figure 5: Closeness centrality rank versus NRC rank by tie inclusion

Given the number of students the most prestigious schools train, excluding student ties should have a significant effect on their centrality. Surprisingly, this is not the case. For the top ten schools, the mean closeness and eigenvector centrality scores using all ties are statistically indistinguishable from those excluding student ties. The top schools do, however, have a statistically higher degree centrality in the graph with all ties than excluding training ties. Thus, while the top schools do train the bulk of professors, they are still at the centre of the later-career hiring net-

work.[10] Another way to show this is illustrated in figure 5 which shows the average predicted rank using closeness centrality for graphs including and excluding student ties. Given over-training, we would expect those graphs excluding student ties to be the better predictors of rank but this is not the case. Instead, the graphs without student ties are often closer to the actual rank. Predictions are better including student ties for Chicago, UCLA, Stanford, and Northwestern while predictions are better excluding student ties for Washington, Harvard, Chapel Hill, Michigan, Berkeley, and Wisconsin. Thus highly ranked schools are central to academic hiring networks whether or not we consider their training role.

We can also test the importance of training using a k-core analysis. First, we separate the graphs into subgraphs where each node has at least degree k within the subnetwork. The subgraphs are calculated by recursively pruning those nodes with degree less than k, producing subnetworks that are interconnected at the same level. This groups together nodes based on both their clustering and their relative popularity, leaving the highest k-core to include the most prestigious departments. However, this changes when training ties are removed. Among the top ten schools 3 appear in the top k-core more often using all ties than excluding student ties. More striking, many more foreign departments enter the top k-core when we exclude training ties (European University Institute, Cambridge, the London School of Economics, and Oxford appear in the top k-cores more than 50% of the time when training ties are removed). Foreign institutions are more important when we remove training but include non-tenure ties because many academics visit the same foreign schools. Excluding non-tenure track positions, foreign institutions do not enter the top k-core at all.

Figure 6 shows the graph excluding PhD training from sample 2, an exceptional graph in the analysis, as it is the only one where the top schools were not ranked highest. In this graph the top k-core was dominated by foreign institutions (LSE, Hebrew University, McGill, University of Quebec, Montreal, and Paris, University of San Paolo, and Oxford) and also included some domestic institutions (NYU and UCSD). Inspecting the raw bipartite graph, it is clear that there are two main clusters, the foreign cluster and the traditional "top" cluster. Both have many prestigious individuals and institutions in them, but one cluster is largely foreign and slightly larger than the second group of traditionally prestigious schools. In sum, the k-core analysis shows that the top ten schools lose some of their dominance without training ties, and that the top British institutions are a central part of the American sociology labor market.

[10] The same is true using top 20 schools.

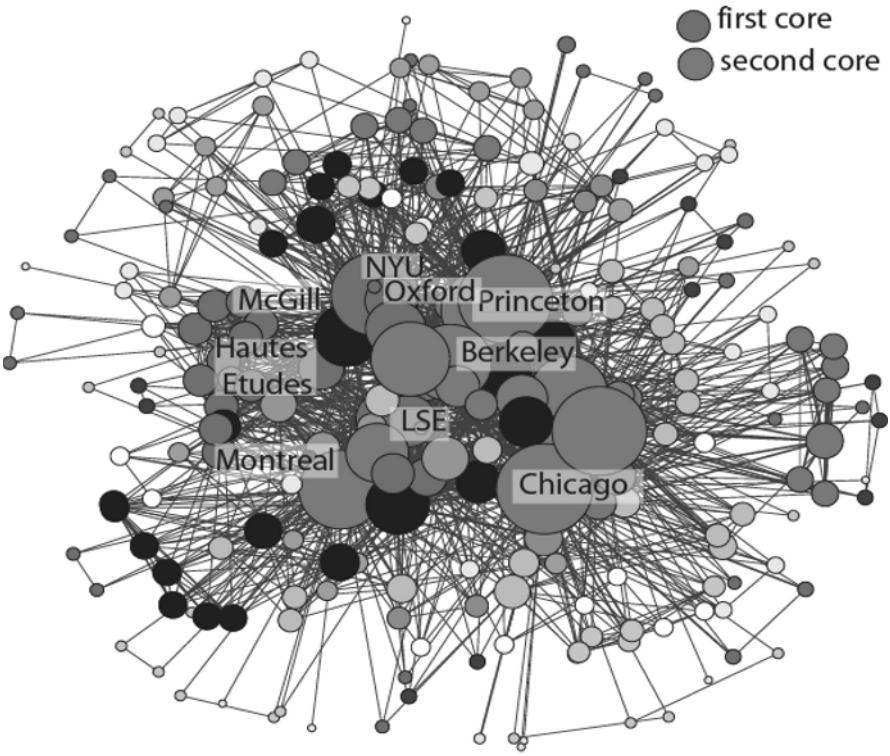


Figure 6: K-core for the reduced sample 2 graph excluding PhD training ties

I test the hypothesis of whether department size matters by first running bivariate regression between each of the centrality measure ranks and the actual academic rank. Then faculty size and the variables related to graph specification are added in, showing that faculty size accounts for very little of the relationship between hiring network centrality and academic ranks. Finally, all the centrality scores are used as predictors in the same equation followed by a Wald test of equality between the centrality measures' coefficients. Results are illustrated in table 5.

Each of the three centrality scores has approximately the same impact regardless of whether or not we consider faculty size. A one position increase in centrality rank predicts at least a .5 position increase in NRC or Newsweek academic rank. Running the three centrality measurements together, we see that eigenvector centrality provides no information not provided by the other two measures and in

fact, controlling for the other two factors, has a negative effect. The only variable related to graph specification that predicts prestige is whether or not the graph includes PhD training ties. When the graph includes training ties, the average predicted ranks increase, because centrality scores are weighted to the academic institutions in the regression analysis and away from the non-academic institutions not included in the regression, because they have no academic rank (such as the census bureau).

Table 5: Predicting prestige with OLS regressions

	domestic rank				international rank			
eigenvector	.481***			-.0367	.588***			-.196*
centrality	**.542****				**.608****			
closeness		.525***		.339***		.666***		.437**
centrality		**.584****				**.667****		
degree			.516***	.241***			.670***	.436***
centrality			.577***				**.686****	
faculty size	-.492***	.460***	-.499***	-.466***	-.121	-.077	-.067	-.063
all edges	.260	.477	.228	.395	.006	.307	.265	.373
no student edges	2.33*	2.47-	1.86	2.29*	-.891	-.629	-.13	-.202
bipartite	.120	.089	.464	.265	.0006	.248	.097	.225
sample 1	.461	.461	.954	.650	-1.27	-1.51	-1.85	-1.73
R-square	.514	.571	.560	.583	.253	.312	.319	.330
coefficient tests	ßeig=ßdegree P: .0013				ßeig=ßdegree P: .00101			
	ßeig=ßcloseness P: .0004				ßeig=ßcloseness P: .0036			
	ßcloseness=ßdegree P: .0004				ßcloseness=ßdegree P: .9789			

bold text indicates bivariate regressions
*** indicates significance at the .001 level

Finally, going back to our first hypothesis, that the prestigious schools maintain their positions by overtraining, we find that running the same regression for only the top 50 schools in the sample and excluding PhD training ties increases predicted rank at least 2 points. This is the opposite of what one might expect if the top schools over-train and rely on placing fresh PhD students to increase their standing in the field. Further, running these same regressions for all observations,

still using only those graphs excluding PhD training ties, a one position increase in eigenvector rank is still correlated with a .42 increase in domestic academic rank and a one point increase in closeness centrality rank is related to a .45 increase in prestige. In sum, excluding student relationships slightly weakens the relationship between graph centrality and prestige, but overall the relationship is still strong.[11]

4 Conclusion

This paper began with two main hypotheses regarding the relationship between the sociology academic employment network and academic rankings. First, we suggested that the relationships might be driven by a spurious relationship with department size. Second, we posited that the relationship could entirely be driven by the dominance of a few departments training the bulk of sociologists and the over-training of sociologists.

I found support that both of these hypotheses are true. Faculty size does explain some of the relationship between centrality and academic rank, and the top institutions are somewhat less central when we consider their dominance in training new PhDs, and that the relationship between centrality and rank is somewhat weaker when we exclude PhD training ties from the analysis. That said, the positive support for these two alternative hypotheses in no way diminished the strength of the relationship between academic rank and centrality in the academic hiring network.

Other researchers finding similar patterns interpret this as an academic "caste system" or infer that training and placement consolidate departments' prestige (Burris, 2004). While I find evidence confirming these patterns, I hesitate to consider it a "caste system" per se and perhaps would consider it a case of positive feedback. If faculty moved strictly in castes (prestigious faculty moving between prestigious institutions and other faculty moving among the other institutions) we would not see this strong relationship between hiring network centrality and academic rank. Rather, we would see two separate cores, lower ranked schools trading with each other and higher ranked school trading with each other. Instead peripheral schools trade faculty with the most prestigious schools rather than with each other. They do this first by hiring graduates of the more prestigious schools,

[11] A Wald test of equality between the centrality scores' coefficients indicates that for both domestic and foreign rank the effects of eigenvector centrality is significantly different from both closeness and degree, though closeness and degrees' effects are statistically indistinguishable from each other.

and then by passing their successful professors on to the more prestigious schools. It is these trades that keep the most prestigious schools in the centre of the graph when we exclude training ties and it is possibly a consequence of this process that highly ranked universities remain highly ranked.

Future work should use data from a wider sample of departments. With a larger sample one might also pursue block modeling, testing whether most ties occur within two distinct groups of departments, as the caste hypothesis should imply. A second possible improvement would include recoding the data as sequential cohort level data. With that sort of data we might test whether the network has become more stratified over time and we might test whether prestigious institutions consolidate their advantage by hiring more accomplished faculty later in their careers. Third, it would be interesting to develop our own measure of research quality, and control for this in predicting departments' prestige. In addition, a study of the evolution of the network would be particularly interesting- testing whether the relationship between trading faculty and academic prestige has been constant over time. Finally, with 40% of new positions in Academic Sociology being adjunct positions, perhaps one of the most pressing questions is what role those adjunct ties play in the network (American Sociological Association, 2007).

References

American Sociological Association. (2007). Profession Trend Data.

Bair, J. (2003). Hiring Practices in Finance Education. *American Journal of Economics and Sociology*, 62, 429–433.

Baldi, S. (1995). Prestige Determinants of First Academic Job for New Sociology PhDs 1985-1992. *The Sociological Quarterly,* 36, 777–789.

Bonacich, P. (1972). Factoring & Weighting Approaches to Status Scores & Clique Identification. *Journal of Mathematical Sociology*, 2,113–120.

Borgatti, S., & Everett, M. (1997). Network Analysis of 2-mode Data. *Social Networks*, 19, 243–269.

Burke, D.L. (1988). *The New Academic Marketplace*. Greenwood Press.

Burris, V. (2004). The Academic Caste System: Prestige Hierarchies in PhD Exchange Networks. *American Sociological Review*, 69, 239.

Davis, K. & Moore, W.E. (1945). Some Principles of Stratification. *American Sociological Review*, 10, 242–49.

Ehrenberg, R.G. (2002). Reaching for the Brass Ring: The US News and World Report Rankings and Competition. *The Review of Higher Education*, 26, 145-62.

Fowler, J.H., Grofman, B., & Masuoka, N. (2007). *Social Networks in Political Science: Hiring and Placement of PhDs, 1960-2002*. Technical report.

Faust, K. (1997). Centrality in Affiliation Networks. *Social Networks*, 19, 157-191.

Graham, H.D., & Diamond, N. (1997). *The Rise of American Research Universities: Elites and Challengers in the Postwar Era*. Johns Hopkins University Press.

Grannis, R. (2005). Sampling the Structure of Large-Scale Social Networks. Technical report, *Proceedings of the Sixth International Conference on Social Science Methodology:* Recent Developments and Applications in Social Research Methodology.

Hargens, L.L., & Hagstrom, W.O. (1966). Sponsored & Contest Mobility of American Academic Scientists. *Sociology of Education*, 39, 24-38.

Hughes, R.M. (1925). *A Study of the Graduate Schools in America.* Miami University Press.

Kleinberg, J.M. (1998). Authoritative Sources in a Hyperlinked Environment. Technical report, *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms*.

Kossinets, G. (2006). Effects of Missing Data in Social Networks. *Social Networks*, 28, 247–268.

Lempel, R., & Moran, S. (2000). The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks*, 33, 387–401.

Lipset, S.M., Bendix, R. & Malm, T.F. (1955). Job Plans & Entry into the Labor Market. *Social Forces*, 33,224–232.

McGinnis, R., & Long, J.S. (1988). Academic Labor Markets & Careers, chapter Entry into Academia: Effects of Stratification, *Geography and Ecology*, pp. 52–73. Palmer Press.

National Research Council. (1982, 1995). *Research-Doctorate Programs in the United States. National Academy of Sciences*.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The Page Rank Citation Ranking: Bringing Order to the Web*. Technical Report SIDL-WP-1999-0120, Stanford University.

Reskin, B. (1979). Academic Sponsorship & Scientists' Careers. *Sociology of Education*, 52, 129–46.

Robins, G., & Alexander, M. (2003). Small Worlds Among Interlocking Directors: Network Structure & Distance in Bipartite Graphs. *Computational and Mathematical Organizational Theory*, 10, 69–94.

Smelser, N.J., & Content, R. (1980). *The Changing Academic Market: General Trends & a Berkeley Case Study*. University of California Press.

US News and World Report. (1995, 2005). *America's Best Graduate Schools*. US News and World Report.

Wiggins, A., Adamic, L., & McQuaid, M. (2006). The Small Worlds of Academic Hiring Networks. *Technical report, Applications of Social Network Analysis Conference.*

A QAP Network Analysis of Intergovernmental Cooperation between Swiss Cantons

Daniel Bochsler

Abstract Swiss cantonal governments cooperate intensively with each other. Inter-governmental treaties (concordats) are a crucial element of the Swiss federalism, but, the intensity of collaboration varies widely. Based on a two-step regression model, this article aims at explaining the different intensities of cooperation. A 'Quadratic assignment procedure regression' (QAP), identifies relational factors among pairs of cantons (common language, political proximity, neighbourhood), whereas an OLS regression, based on attribute data, explains why different cantons (nodes) have different strong motivations for collaboration.

Acknowledgments I am very grateful to Alex Fischer, Manuel Fischer, Pascal Sciarini, Balazs Vedres, two anonymous reviewers and the volume editor for comments and help and to Samuel Thomi for graphical assistance.

1 Introduction

In the Swiss political system, the cantonal (sub-national) governments have far-reaching powers in many relevant policy fields. Switzerland includes twenty-six different sized cantons, wit populations ranging from 10,000 to over 1.2 million residents. When the Swiss federal state was founded in 1848, the federal government obtained only minimal powers (namely customs, postal services, currency, defence, foreign affairs). Even if these have increased in importance over time, the Swiss cantons still control very important political fields, such as justice and police, health services, elementary, secondary education, and large parts of tertiary education (Vatter, 2007). Despite strong economic, cultural and social ties across cantonal borders, and despite an increase of inter-cantonal ties and mobility over

time, the cantonal structure has remained hardly touched at all in the latest 160 years of history.

In recent decades, cantonal governments have intensified their cooperation. Such cooperation is seen as a way to hold powers at a low level of the state, and being able to coordinate policies which address a more and more nationalised society (Bochsler/Sciarini, 2006a). The cooperation of the cantons is based on so-called concordats (inter-governmental treaties at the cantonal level). Typical bilateral or multi-lateral concordats address topics such as the access of cantonal schools by students of other cantons, inter-cantonal institutions in the penal system, mutual support of cantonal police forces from different cantons in the case of events which exceed the capacity of a single cantonal police, or inter-cantonal rules about the exception from (cantonal) taxes on inheritance. Today, at least some 760 concordats exist, each with 2 up to 26 affiliated cantons. Each canton is related with each other through at least a 16 concordats, but the intensity of collaboration varies widely: Some pairs of cantons count even more than 100 common concordats (Bochsler/Sciarini, 2006a).

The cooperation of Swiss cantonal governments is more pronounced than in other federal states. While Bolleyer (2006) has related this to the consociational model of government, this study tests for alternative explanations that might account for the varying intensity of cooperation. My model relies on the costs and benefits from inter-cantonal cooperation: neighbouring cantons and cantons that share a common language might have more benefits from cooperation, while political differences make cooperation more costly. Arguments linked to economy of scale make us expect that small cantons might be more inclined to agree to concordats in order to implement policies in a cooperative way.

The structure of inter-cantonal cooperation has been analysed with tools from network analysis. Bochsler and Sciarini (2006a) have mapped the pattern of inter-cantonal cooperation with the multidimensional scaling technique, but a systematic quantitative analysis of the intensity of cooperation is so far lacking.

The network of cooperation of Swiss cantons appears as an appropriate case for the application of analytical instruments from network analysis. This article employs a two-step regression model. In the first step, I use QAP regression analysis in order to explain the varying intensity of network contacts among the Swiss cantons. This model accounts for the number of contacts for each of the 325 (26*25/2) pairs of cantons (relational data). At the second level of my analysis, I explain different levels of cooperation for each of the twenty-six cantons (attribute data).

This article proceeds as follows. First, a review of previous research shows the lack of a quantitative model applied on the research question. After having shown that there is substantial variance in inter-cantonal cooperation, I introduce my own model which explains incentives and disincentives for Swiss cantons to join intergovernmental treaties. This model is tested then, using a database comprising the concordats between the Swiss cantons, and using relational data on common characteristics of pairs of cantons.

2 Previous research

The concordats between Swiss cantons are an important pillar of the Swiss horizontal federalism. The term "horizontal federalism" is employed for processes of collaboration between institutions at the same level of the state – such as the cooperation between the Swiss cantons, not involving the upper level. Horizontal federalism is different from vertical cooperation, where different levels of state cooperate with each other, for instance when the federal state cooperates with its subnational units. The first concordats that are still in use date from the founding years of the Swiss federalism or even before. However, most of the concordats have been established after 1965, when the Swiss system of "horizontal federalism" gained importance.

This is reflected too in an increasing interest of research for the cooperation among Swiss cantonal governments. Most studies are interested in juridical aspects, such as the problem of democratic legitimacy of the new meta-level of governance (Abderhalden 1999; Boegli 1999; Schöni 2005: 17-8; Rhinow 2002; Kramer 1997: 282). Others focus on the impact of the internationalisation of politics, foremost caused through the current process of integration of Switzerland into the European institutions, on the role of cantons in the Swiss federal structure and the reinforcement of horizontal cooperation (Häusermann 2003; Abderhalden 2000; Minger 2004).

Quantitative analytic studies of the system of Swiss horizontal federalism have remained rare. Bochsler and Sciarini (2006a) have used network analysis tools (multidimensional scaling) to map the pattern of inter-cantonal cooperation. Their study is based on a database which lists some 760 treaties among Swiss cantons, most of them with just 2 cantons, but others counting more members. With some of the treaties, all 26 cantons acceded. The study recognises patterns of regional cooperation: often, it seems, neighbouring cantons are collaborating most closely with each other. Furthermore, four small groups of cantons have a large number of

common treaties (French- and Italian-speaking group of cantons; Northwest; Central; Eastern Switzerland).

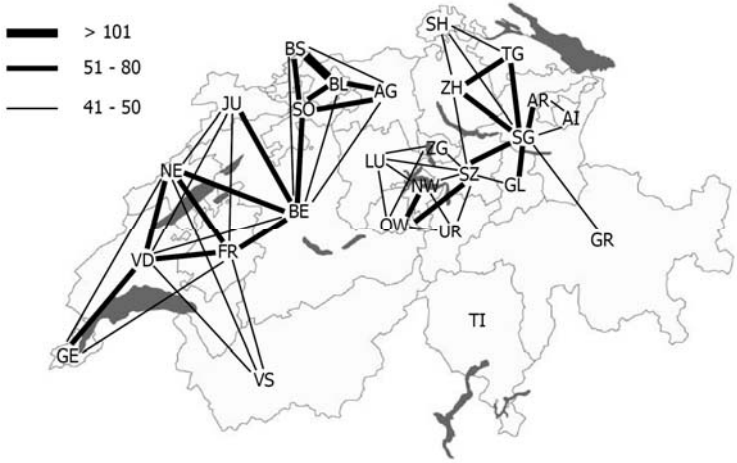


Figure 1: Map of the Swiss cantons and their common concordats
Only connections of 41 or more concordats shown. See list of abbreviations of Swiss cantons in the appendix. (Map drawn by Samuel Thomi)

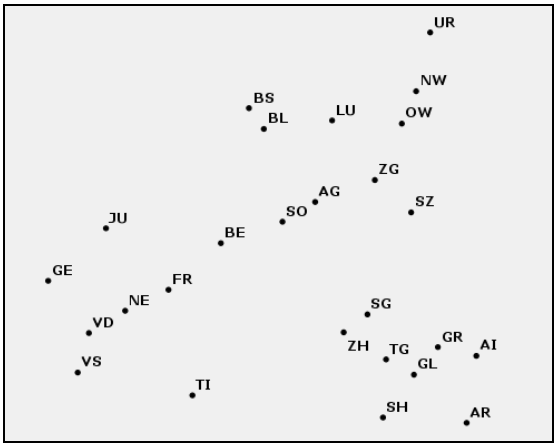


Figure 2: Structure of the system of concordats among Swiss cantons
See list of abbreviations of Swiss cantons in the appendix. (Figure taken from Bochsler/Sciarini, 2006a. Sources: Database University of Fribourg/database Bochsler/Sciarini, for concordats existing in 2005)

Bolleyer (2006) has looked at reasons why Switzerland has developed such strong cooperation among cantons, stronger than other federal states in the international comparison. The main argument put forward to explain this enhanced cooperation is based on the fact that Swiss federal and cantonal governments often resemble all-party-coalitions, with a quasi-proportional division of the government posts on all parties (see Bochsler/Sciarini, 2006b; Vatter 2007, 204). Such all-party coalitions might simplify intensive cooperation among cantons, compared to other federal countries. In countries where sub-national governments have the form of minimal-winning coalitions, one or few parties are represented in the cabinet. In consequence, governments of two different sub-national entities often belong to different parties, and might perceive each other as political enemies. This might be an obstacle for their cooperation. Not so in the case of the Swiss cantons: if almost all major parties are represented in each cantonal government, most cantonal governments look similar, and there are less political obstacles which might hinder their cooperation (Bolleyer, 2006).

However, there is no known systematic quantitative test including competing hypotheses which might explain the intensity and the structure of the system of concordats. This research gap is filled with this article.

3 Differences in inter-cantonal cooperation

My analysis is built on possibly the most complete inventory of concordats, which has been built up by the University of Fribourg, and completed and adopted for quantitative analyses by Bochsler et al. (2004, 94-9) and Bochsler/Sciarini (2006a). It counts about 760 concordats that were listed in different sources.[1] A first, descriptive analysis shows the number of concordats by cantons. The results reported in figure 2 show that some of the cantons are much more involved in inter-governmental cooperation than others. With its 220 titles, the canton of St-Gall is leading the list. The canton of Ticino has only forty concordats, the lowest number of all cantons. What are the reasons for this lack of collaboration? Two

[1] Previous databases were built up by Frenkel/Blaser (1981), the Zentralschweizer Regierungskonferenz (only listing concordats among cantons of Central Switzerland, http://www.zrk.ch /prog/default.asp?struktur_id=57), and the University of Fribourg (http://federalism.unifr.ch /concordat/ge/index.html). It is plausible that the database is not complete. There might be some treaties which are not reported in the listings of the University of Fribourg on which the analyses are based. For the present analyses, there is however little reason why possible gaps would be so systematic that they would lead to biased results.

explanations appear plausible. The first is Ticino's territorial isolation in the South of Switzerland. From the rest of the country, Ticino is separated by the Alps, and for some relations transit through neighbouring Italy is needed to reach the canton. The second is the linguistic isolation: Ticino is the only canton with an Italian-speaking majority of the population, and only one other canton has an (autochthonous) Italian-speaking minority.[2] The regression model later in this article might clarify the puzzle.

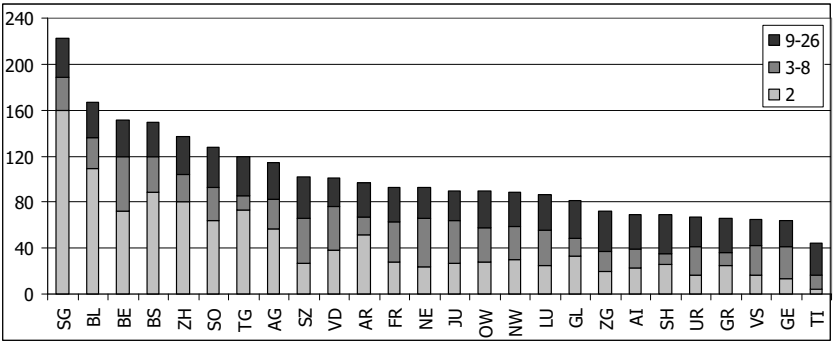


Figure 3: Number of concordats by canton and size of the concordat (by the number of member cantons). See list of abbreviations of Swiss cantons in the appendix

My analysis not only investigates the density of concordats by canton,[3] but also the network structure of concordats between the twenty-six cantons. For this purpose, I built a relational variable, counting the number of concordats for each pair of cantons, listed in a matrix of 26 x 26 cantons. Since the network structure is symmetric, there are 325 cases (or pairs of cantons). The two neighbouring cantons Basle-City (BS) and Basle-Country (BL) are best connected to each other, counting 119 common concordats. The least connected ones are Valais (VS) and Appenzell Outer-Rhodes (AR), two cantons which are geographically very distant from each other, with sixteen network ties. On the average, each Swiss cantons is cooperating with each other in thirty-one concordats. Further, I differentiate concordats by six policy fields, and test if the explanation model is universally applicable for all of them.

[2] In other areas, there is a small percentage of Italian speakers, but they are mainly recent migrants from Italy, so that they do not have the status of a traditional language group.

[3] This would only indirectly reflect the structure of the network of concordats, and rather give us information about the centrality of a node in the network.

4 Explaining cooperation between cantons

Because the main focus of this article is on methodological aspects, I introduce my explanatory model for the varying intensity of cooperation only shortly (for a closer discussion of this model, see Bochsler/Sciarini, 2006a). Based on several explanatory factors, I formulate testable hypotheses for the statistical analysis.

Hypothesis 1: Smaller cantons cooperate more often.

Cantonal cooperation has been discussed as a consequence of the size of Swiss cantons. Compared to federal units in other countries, they are very small, often too small to implement policies in an efficient way (Sciarini, 2005). The smaller a canton is, the more it should thus be interested in finding a partner for cooperation, in order to decrease its costs of policy implementation, either with another small canton or with a large canton. This leads to the expectation that network ties are particularly strong if one of the cantons is small. To operationalise this hypothesis, for each pair of cantons, I considered the number of inhabitants of the smaller of both cantons.[4]

Hypothesis 2: Neighbouring cantons cooperate more strongly.

In many fields, cooperation is only useful or even possible if two cantons are located close to each other. Some typical subjects of inter-cantonal treaties might underline this point: In health policies, the geographical position of a hospital matters. For street maintenance, the existence of inter-cantonal streets is crucial. Concordats about fishing are agreed on when rivers or lakes cross cantonal borders. The geographical proximity hypothesis applies particularly for territorially-based policies.[5] There are two ways that proximity might be operationalised: First, through the existence of a common border between two cantons (dummy variable), because in some policy fields, only cooperation between direct neighbours makes sense. Second, I include a measure of the geographical distance of two cantons (the distance between their capitals). This second measure has the advantage of being metrically scaled, and it reflects that two cantons which are located close to each other, even if not immediate neighbours, might have an interest in cooperating, or they might belong to a regional network of intense collaboration.

Hypothesis 3: Language matters for cooperation.

[4] A different operationalisation, using the mean population of both cantons, was tested, but shows not to lead to better results.

[5] For a discussion of territory and non-territory-related policies and federalsim, see Braun (2000).

In other, non-territory related policy fields, other criterions might be important. In Switzerland as a *pluri-linguistic* country, languages are important for many policies. The country has four official languages, and each linguistic group is territorially concentrated. As a result, most of the cantons have only one official language, and just four out of twenty-six cantons count two or three languages. Cantons with the same language might collaborate more closely, particularly in the education and cultural sector. This idea is expressed through a relational dummy variable, which measures the common use of French, German, or Italian as official language in two cantons of a pair. (Since only one canton uses the fourth Swiss national language, Romansh, officially, it can not be used as relational variable.)

Still, there are policy fields where coordination is neither linked to language, nor to territory (a typical example is the mutual recognition of diplomas which are issued by cantons, cf. Freiburghaus/Zehnder, 2003, 5). This is why coordination occurs as well among distant cantons with different languages, but since there are fewer relevant issues, it might be less intensive.

Hypothesis 4: Political differences are hampering cooperation.

Cooperation might impose *political costs*: namely, there are transaction costs to agree in a concordat, and cooperating cantons might lose a part of their political autonomy. Such costs might be more important if governments with different political views negotiate with each other. In such cases, a common implementation of a policy might require the partners to change their policies. This is why the loss of autonomy might be more significant in the case of political differences (Cameron, 2001; Bolleyer, 2006). As a consequence, cantons with differently composed governments might agree less often on treaties, than such with a similar political colour. I measure political differences with Gallagher's (1991) *Least-Square Index* (for the government composition of 2005, data taken from Bochsler/Sciarini, 2006b). For governments that are identically composed, the index takes the value 0; if two governments are solely composed by members of two different parties, the index is coded 1 (see table 1). I expect that the stronger differences between two governments, the less they cooperate, but the intensity of the effect might depend on the nature of the concordats.

Hypothesis 5: Decreasing marginal utility of cooperation.

Finally, given that cantons agree to concordats because this makes public action more efficient, I expect a decreasing marginal utility of cooperation. Cantons collaborate with others in order to seek positive scale effects. This is why they have substantial interest in finding partners for cooperation. Accordingly, a canton

which can easily establish cooperation in a policy field might be less interested in finding additional partners for cooperation and agree to new concordats than cantons which have difficulties in finding partners for cooperation. Based on this consideration, I expect that if all relational variables are accounted for, pairs of cantons with a low degree of cooperation will rather agree in concordats pairs of well-included cantons. The hypothesis of decreasing marginal utility can not be operationalised directly. I argue that cantons which are surrounded by many potential partners for cooperation are better included in the network of concordats. They will, ceteris paribus, less frequently join a concordat than cantons with an isolated position in the network. The hypothesis can only be investigated in the second step of the regression analysis to follow.

Table 1: Variables included in regression models

name of variable	minimum	maximum	arith. mean	std. dev.
Dependent variable: Number of concordats (relational variables, N=325)				
all policy fields (log)	2.77	4.78	3.38	0.32
education, science, culture	3	31	9.15	4.81
health services, social security	2	21	3.30	1.90
security, state organisation	5	32	9.65	2.61
infrastructure, environment, traffic	0	19	1.56	2.38
economy, agriculture	2	14	4.47	1.71
finances, taxes	1	5	2.67	1.04
Independent variables in the first-step model (relational variables, N=325)				
distance (log)	2.16	5.641	4.472	0.643
common border	0	1	0.16	0.37
partisan differences	0	0.77	0.34	0.13
French	0	1	0.065	0.246
German	0	1	0.646	0.479
Italian	0	1	0.003	0.055
Population of the smaller canton (log)	10.0	13.9	12.0	0.8
Independent variables in the second-step model (variables by cantons, N=26)				
Population (log)	10.0	14.0	12.0	1.1
French speaking	0	1	0.269	0.452

Table 1 gives an overview over the descriptive statistics of the variables which are used to operationalise the explanatory model. The dependent variable and the independent variables in the first-step model are all relational variables, whereas in the second-step model, two exogeneous non-relational variables will be introduced.

5 Regression models

For my quantitative analysis, I proceed in two steps, first analysing relational, and then non-relational variables. My first regression model controls for the four relational hypotheses (1-4). The dependent variable (number of common concordats for a pair of cantons) and some of the independent variables (distance between cantons, population of the smaller canton) are positive, open-scale numbers. I take the logarithm of these variables, because this fulfils the distributional assumptions of the linear regression model better. Further, for each of the twenty-six cantons, a dummy variable is included as an independent variable. With this dummy variable, I consider that some cantons might join systematically more concordats than others, if all other conditions are equal. I assume that cantons have certain characteristics which make them be more or less inclined to join concordats. In a second step, a further regression model explains these differences among cantons.

There are several methods which have been developed for quantitative multivariate analyses of network data. Common multivariate models (OLS regression, logit, etc.) face – when applied on network data – the problem that several independence assumptions are not met by network data. One of the specific network analysis methodology, p*, controls for these dependency effects, and helps to explain the absence or the (symmetric or asymmetric) occurrence of ties between nodes of a network, using relational characteristics and characteristics of the whole network. The method provides *logit models* and explains dichotomous outcomes (Wasserman/Pattison, 1996; Pattison/Wasserman, 1999; Anderson et al., 1999). In my present application, all the nodes are connected with each other symmetrically, so that there is no variation with regards to the existence of ties. Instead, the number of ties (or: the intensity of ties) differs from case to case, and is in the focus of my study. My dependent variable has thus a metric form. *Quadratic assignment procedures* (QAP regression) allows not only the analysis of models where both the dependent and the independent variables are ratio-scaled, it furthermore allows the inclusion of independent variables in the form of network data or to measure similarities and differences among the units which form the

Table 2: Results of the QAP regression model; N=325 pairs of cantons

	(G) all concordats (log)		(1) education, science, culture		(2) health services, social security		(3) security, state organisation		(4) infrastructure, environment, traffic		(5) economy, agriculture		(6) finances, taxes	
	coeff.	stand	coeff.	stand	coeff.	stand	coeff.	stand	coeff.	stand	coeff.	stand	coeff.	stand
constant	1.49	.000	11.51	.000	5.58	.000	1.70	.000	1.79	.000	8.290	.000	4.12	.000
Population (log, smaller canton)	.060**	.182	2.45**	.214	1.20*	.265	.784(*)	.126	-.282	-.050	.102	.025	-.101	-.041
Common border	.046**	.124	1.61**	.124	.243	.047	1.13**	.160	2.04**	.317	1.16**	.252	-.205*	-.073
Distance (log)	-.325**	-.658	-9.88**	-.574	-4.68**	-.687	-4.33**	-.463	-4.75**	-.558	-3.18**	-.519	-.875**	-.235
partisan differences	-.06(*)	-.058	-.844	-.022	-.841	-.056	-.663	-.032	-1.47(*)	-.079	.162	.012	-.379	-.046
Official language														
French	.164**	.294	7.43**	.380	-.417	-.054	2.17**	.205	-1.74**	-.179	.140	.020	1.88**	.444
German	.128**	.445	3.90**	.388	.452	.114	2.69**	.493	.016	.003	.511(*)	.143	.236	.108
Italian	-.120*	-.048	-6.22**	-.072	-1.27	-.037	-1.77	-.038	-1.47	-.034	-1.84*	-.060	.634	.034
Dummy var.	Not reported (available upon request); shall be included in the second step regression model.													
Adjusted R^2	.823		.794		.575		.626		.602		.668		.714	

(*) significant at p < 0.1; * significant at p < 0.05; ** significant at p < 0.01

nodes. It is best suited for my analysis. My dataset consists of a square matrix of 26 x 26 nodes (cantons), counting 325 possible relationships among the nodes. In this matrix, the error terms of the 325 cases are not statistically independent from each other. This problem of lacking independency creates autocorrelation problems, which lead to an underestimation of standard errors. QAP regression models control for autocorrelation. Similar to other regression techniques, QAP calculates the regression coefficients. However, to establish the probability of a factor, it runs real data, but then randomly permutes lines and rows of the dependent variable, which later is compared to the real data (Krackhardt, 1988).[6]

The results of the regression model are presented in table 2. The first model (G) shows the results for the general model, including all policy fields. Further six models are calculated for concordats of specific policy fields. Three out of four hypotheses which are tested in the first-step regression model can be regarded as confirmed. *Geographic proximity* (hypothesis 2) contributes most to the explanation: The variables (*distance* and *common border*) are statistically significant in almost all the models. Interestingly, in the policy field where territoriality plays the least role (*finances, taxes*), the contribution of the variables is lowest, and the common border variable goes in the opposite direction to that expected. In the policy field which is closest linked to the territory (*infrastructure, environment, traffic*), common borders are more important for the explanation than in other fields. In the policy field of *health services and social security*, the geographic distance is particularly important, and common borders irrelevant. This is not very astonishing, because cooperation in health and social services rather relies on the distance than on common borders.*Partisan differences* (hypothesis 4) are not very central for the explanation of inter-cantonal cooperation. As expected, the variable is negatively correlated to cooperation in almost all models. However, in the general model it reaches only a low level of statistical significance, and in the partial models it remains mainly non-significant. And finally, *common official languages* (hypothesis 3) are an important explanation for the density of the networks of concordats: *French speaking* and *German speaking* pairs of cantons have substantially more concordats than pairs of cantons of the reference category (no common official language).[7] As expected, the common language is particularly relevant in the policy field where language plays a major role – *education, science, culture*, while in other fields it reaches only occasional statistical significance (and, in policy

[6] See Burris (2005) for an example. I estimated the regression model with UCINET, using 10,000 permutations of the matrix.

[7] Only two cantons have Italian as a common official language; this variable relies thus only on one positively coded tie, and should not be over-interpreted.

field 4 it appears even relevant in the non-expected direction). The model does not confirm my first hypothesis however, regarding the *size of the cantons*: Small cantons seem not to be more active in cooperation.

In the first-step regression model, the focus was on relational variables, which characterise the common or different features of each pair of cantons and might incite or impede cooperation. Non-relational characteristics of cantons were included through dummy variables for each canton. These dummy variables were measuring the number of concordats a canton joins, compared to its potential which is expressed through the relational variables. The dummy variables are analysed in the second-step regression model. Before, however, I discuss the results of the first-step model, regarding the dummy variables, briefly. Based on the dummy variables estimated in the first-step regression analysis (general model for all policy-fields), it looks like if both peripheral cantons Ticino (TI) and Geneva (GE) were top-rated, with coefficients of 0.23/0.20 (see figure 4).

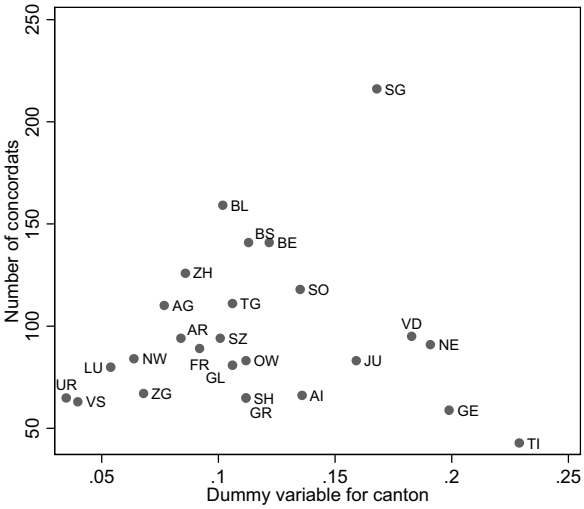


Figure 4: Coefficients from the regression model (table 2) and number of concordats by canton. See list of abbreviations of Swiss cantons in the appendix.

These high values signify that Ticino and Geneva establish more concordats with potential partner cantons than other cantons do, provided that all other relational variables (language, distance to partner canton, confession) are equal. Nevertheless, they are among the cantons with the lowest rate of cooperation. This looks

inconsistent, but it is not. The paradox seems to be an *effect of isolation* or of *decreasing marginal utility*. Geneva and Ticino are geographically isolated at the very Western or Southern end of Switzerland, so that they have few natural partners for cooperation. Adding to this, Ticino is linguistically isolated. This is why both cantons are not well included in the network of cooperation. And, even to reach a low level of cooperation, cantons such as Geneva and Ticino must actively seek cooperation. Or put differently, in situations where other pairs of cantons would not necessarily cooperate with other cantons, Geneva and Ticino do.

On the other hand, cantons such as Basle-Country (BL), Basle-City (BS), or Bern (BE), are amongst the less active cooperation seekers, but have very high numbers of concordats. This is due to their geographic proximity to other cantons, and, in the case of Bern, due to its two official languages (which make it an attractive partner for cooperation both with German *and* with French speaking cantons). Overall, figure 4 does however not show a clear correlation. Nevertheless, some cantons which are peripheral in the network of cooperation would agree on many concordats, given their low network potential.

Such phenomenon corresponds to the expectation of the *decreasing marginal utility*: The more concordats a canton might establish due to its natural network potential, the lower its utility of any additional concordat, so that it will be a less active cooperation seeker. Cantons with few concordats (isolated cantons) are highly interested in finding partners for cooperation, because due to the small size of Swiss cantons it is very costly not to cooperate. This is why cantons with few concordats cooperate even in the case of an unfavourable context (e.g. with the political opponent, with a canton of a different language, and over big distances).

To measure this hypothesis more systematically, I operationalise it with two measures which both derive from the first-step regression model. The dependent variable are the coefficients for the dummy variables which I have estimated in the first-step regression model. They measure whether a canton joins more concordats than other cantons would do, given their natural network potential, or, under same (geographical, linguistic, political) conditions, how often a canton joins concordats with a potential partner.

I expect that these coefficients depend on the potential of a canton to establish a network of cooperation. It would be problematic however to measure this through the number of concordats which a canton has established – this measure is endogeneous to the dummy variable. This is why, instead of the real number of network ties of cantons, I measure the potential of the cantons to establish concordats, which I call the *natural network potential* of a canton. This is the number of concordats that a canton would have, due to its geographical location, its lan-

guage, and due to the political colour of its government. Based on the first-step regression model, I can estimate for every combination of cantons the number of concordats that would be established if under similar conditions, every canton would be similarly likely to join a concordat.[8] For every canton, I estimate the *natural network potential* through the addition of the potential number of links with each of the 25 other cantons. Isolated cantons have a low natural network potential, while cantons with many potential partners for a close cooperation have a high *network potential*. Typically, centrally located cantons, with many potential partners that share the same official language and a similar partisan composition of the government have a high network potential. Cantons that are geographically, politically, and linguistically isolated have a low potential for cooperation. The official language of the cantons (dummy for French language) and their size are included as control variables.[9]

The OLS regression models (table 3) confirm the hypothesis of the decreasing marginal utility (hypothesis 5). The first model regards all concordats, and six further models by policy field are presented. The *natural network potential* is the most important in six out of seven models, and statistically significant. As expected, cantons with *lower* network potential rather join concordats (their dummy variables have higher coefficients). This shows that it is essential to cantons with few (potential) concordats to find cooperation partners. They will even involve in concordats if the relational variables might make it difficult (or less beneficial) to cooperate. For cantons with few concordats, it appears more important to reach additional concordats than for such that have already have established some cooperation.

Further, the size of the cantons (population) seems to have an impact in several models, although not always statistically significant. As in the first-step regression models, the direction of the coefficient is however opposed to the expected direction, with large cantons cooperating more often.

[8] Since the dummy variables are not exogeneous to the model, they are set at average (geometrical mean of all the dummy coefficients).

[9] The dependent variable is not a naturally measured variable, but one derived from another regression model. This makes it particularly important to control for deviations through heteroscedacity (Lewis/Linzer, 2005).

Table 3: Results of the OLS regression model; N=26 cantons (model 4: 24 cantons), robust standard errors

	(G) all concordats (log)		(1) education, science, culture		(2) health services, social security		(3) security, state organisation		(4) infrastructure, environment, traffic		(5) economy, agriculture		(6) finances, taxes	
	coeff.	stand	coeff.	stand	coeff.	stand	coeff.	stand	coeff.	stand	coeff.	stand	coeff.	stand
constant	1.59		20.37		5.83		9.33		-.179		9.42		1.44	
Population (log)	.015**	.337	.227	.146	-.123	-.192	.264(*)	.190	.126	.232	.182	.324	.190(*)	.383
French speaking	-.006	-.051	.350	.090	.380	.237	-.007	-.002	-.008	.309	-.66(*)	-.473	-.49(*)	-.391
Potential for concordats (log)	-.280**	-.834	-4.76**	-.699	-1.10**	-.390	-3.05**	-.841	-.224*	.088	-2.70**	-.577	-.938	-.107
Adjusted R^2	.702		.528		.628		.764		.199		.312		.267	

(*) significant at p < 0.1; * significant at p < 0.05; ** significant at p < 0.01

6 Conclusion

The present analysis is the first quantitative analytic investigation into the motivations of Swiss cantons to join concordats and into the determinants of the network of inter-cantonal cooperation. Since many of the cantonal powers are related to the territory, geographic aspects play an important role for the intensity of cooperation. Neighbouring cantons are most active partners in the network of concordats. Language plays the second most important role: cantons with the same official language have more common treaties than these with different languages. Politics matter only in third instance. It has been hypothesised (Bolleyer, 2006) that the costs for governments to cooperate are substantially higher when they are of different political colour, but the impact is not particularly strong. This might be related to the rather technical character of many of the concordats. Further, it can be observed that the importance of explanatory factors varies among different policy fields. For territorial policies, neighbourhood and geographic proximity of cantons are important for cooperation; for education and cultural policies, language matters much more.

Finally, cantons which are rather isolated, in linguistic, political, and geographic terms, are more active in joining concordats, compared to cantons which are central in the network. This can be described as an effect of decreasing marginal utility: In order to get more efficient, cantons need to cooperate. However, once they established some concordats, the marginal utility of additional concordats decreases. Accordingly, cantons which are very central in the network and can easily establish links to other cantons are less active in seeking cooperation. They already have network links and for them an increase of cooperation is less important than for cantons which are geographically, linguistically and politically isolated.

The analysis has been carried out in two steps, first explaining the reasons why cantons cooperate with each other (relational model; QAP regression). But the analysis is not limited at the application of a QAP model: It shows, how the results of the QAP model can be linked with other quantitative analysis methods – in the case of this article with an OLS regression. In this analysis, this happened through the inclusion of dummy variables for each node in the network in the QAP model, which are further analysed in a second step. Further, the second-step analysis included residuals of the first-step model which allow to measure the centrality of nodes.

The article presents the methodology based on a case-study of the Swiss case. However, the proposed analytical two-step methodology, linking a first relational

model with a secondary, non-relational analysis, might be similarly applied for other studies too.

References

Abderhalden, U. (1999). *Möglichkeiten und Grenzen der interkantonalen Zusammenarbeit*. Universitätsverlag, Freiburg.

Abderhalden, U. (2000). Möglichkeit und Grenzen der interkantonalen Zusammenarbeit bei der internationalen Integration der Schweiz. In P. Hänni (Ed.), *Schweizerischer Föderalismus und europäische Integration: die Rolle der Kantone in einem sich wandelnden internationalen Kontext* (pp. 323-381). Zürich: Schulthess.

Anderson, C.J., Wasserman, S.,& Crouch, B. (1999). A p* primer: logit models for social networks. *Social Networks*, 21(1), 37-66.

Bochsler, D., et al. (2004). *Die Schweizer Kantone unter der Lupe. Behörden – Personal – Finanzen*. Bern: Haupt.

Bochsler, D.,& Sciarini, P. (2006a). Konkordate und Regierungskonferenzen. Standbeine des horizontalen Föderalismus. *Leges*, 2006/1, 23-41.

Bochsler, D.,& Sciarini, P. (2006b). Neue Indikatoren zur Bestimmung der arithmetischen Regierungskonkordanz. *Swiss Political Science Review*, 12(1), 105-122.

Boegli, L. (1999). Les concordats intercantonaux: Quels enjeux pour la démocratie? *IDHEAP, Travaux de cours et mémoires de l'Idheap*, 12/1999.

Bolleyer, N. (2006). Consociationalism and Intergovernmental Relations. Linking Internal and External Power-Sharing in the Swiss Federal Polity. *Swiss Political Science Review*, 12(3), 1-34.

Braun, D. (2000). Territorial Division of Power and Public Policy-Making: An Overview. In D. Braun (Ed.), *Federalism and Public Policy* (pp. 1-26). Ashgate: Aldershot.

Burris, V. (2005). Interlocking Directorates and Political Cohesion among Corporate Elites. *American Journal of Sociology*, 111(1), 249-283.

Cameron, D. (2001). The structures of intergovernmental relations. *International Social Science Journal*, 53(167), 121-127.

Freiburghaus, D.,& Zehnder, V. (2003). Horizontale Kooperation zwischen den Kantonen und die "systematisch-pragmatische Zusammenarbeit" in der Zentralschweiz. *Working paper de l'IDHEAP*, 4/2003.

Frenkel, M.,& Blaser, T. (1981). *Konkordanzregister, Verzeichnis der Ende 1980 geltenden interkantonalen Verträge mit einer Kurzdarstellung des schweizerischen Konkoratsrechts*. Riehen: Institut für Föderalismus und Regionalstrukturen.

Gallagher, M. (1991). Proportionality, Disproportionality and Electoral Systems. *Electoral Studies*, 10(1), 33-51.

Häusermann, S. (2003). Internationalisation des politiques publiques et mise en œuvre fédéraliste – La libéralisation des marchés publics cantonaux en Suisse. *Cahier de l'DI-HEAP*, 209/2003.

Krackhardt, D. (1988). Predicting With Networks: Nonparametric Multiple Regression Analysis of Dyadic Data. *Social Networks*, 10 (4), 359-381.

Kramer, U. (1997). Die Funktion der EDK zwischen gestern und morgen. In: H. Badertscher. (Ed.), *Die Schweizerische Konferenz der kantonalen Erziehungsdirektoren 1897 bis 1997* (pp. 273-292) Bern.

Lewis, J.B.,& Linzer, D.A. (2005). Estimating Regression Models in Which the Dependent Variable Is Based on Estimates. *Political Analysis*, 13, 345-364.

Minger, T. (2004). Die Geschichte der Konferenz der Kantonsregierungen. In: Konferenz der Kantonsregierungen, *10 Jahre Konferenz der Kantonsregierungen 1993-2003. Standortbestimmung und Ausblick*. Bern: KdK.

Pattison, P.,& Wasserman, S. (1999). Logit models and logistic regressions for social networks: II. Multivariate relations. *British Journal of Mathematical and Statistical Psychology*, 52, 169-193.

Rhinow, R. (2002). Wenig autonomie- und demokratieverträglich. Staatspolitische Bedenken zur Ausgestaltung des neuen Finanzausgleichs, *Neue Zürcher Zeitung*, 7. Mai 2002.

Sciarini, P. (2005). Die interkantonale Zusammenarbeit ebnet den Weg für die Föderalismusreform. *Neue Zürcher Zeitung*, 8. Februar 2005.

Schöni, A. (2005). Le contrôle parlementaire des conventions intercantonales. *Working paper de l'IDHEAP*, 4/2005.

Vatter, A. (2007). The Cantons. In: U. Klöti, U. et al. (Eds.), *Handbook of Swiss Politics* (pp. 197-226). Zurich: NZZ.

Wasserman, S.,& Pattison, P. (1996). Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and p*. *Psychometrika*, 61(3), 401-425.

Appendix: Abbreviation of cantons

AG: Aargau; AI: Appenzell Inner-Rhodes (Appenzell Innerrhoden); AR: Appenzell Outer-Rhodes (Appenzell Ausserrhoden); BE: Berne (Bern); BL: Basle-Country (Basel-Landschaft); BS: Basle-City (Basel-Stadt); FR: Fribourg; GE: Geneva (Genève); GL: Glarus; GR: Grisons (Graubünden); JU: Jura; LU: Lucerne (Luzern); NE: Neuchatel; NW: Nidwalden; OW: Obwalden; SG: St. Gall (St. Gallen); SH: Schaffhouse (Schaffhausen); SO: Solothurn; SZ: Schwyz; TG: Thurgau; TI: Ticino; UR: Uri; VS: Valais; VD: Vaud; ZG: Zug; ZH: Zurich (Zürich).

Structural Changes in Agent-Based Simulations: Representing HIV/AIDS Impact on Social Networks

Shah Jamal Alam & Ruth Meyer

Abstract HIV/AIDS in the Sub-Saharan Africa is one of the biggest threats against sustainable human development in the region. One of the implications of this epidemic is that it not only affects individuals and their households but also increases burden on traditional social and support networks or safety nets. Research reported in recent years, have indicated on the possibility of weakening and even breaking of these support networks. In this paper, we suggest that by analyzing dynamical networks generated from agent-based simulations, one could describe this effect with better precision. This idea is based upon our previous work, which attempts at finding techniques to identify structural changes in dynamic networks.

Acknowledgments We would like to thank Thomas Friemel, the editor, for support and useful comments. Also to thank Scott Moss, Bruce Edmonds, Armando Geller, our other colleagues at the Centre for Policy Modelling and the Stockholm Environment Institute (Oxford) for their comments and feedback. This work is supported under the EU FP6 Project CAVES.

1 Introduction

AIDS in Sub-Saharan Africa is a serious problem, as it not only threatens lives of individuals susceptible to the infection but also the social structure of a community itself. In many parts of the rural Sub-Sahara, where employment opportunities are few, majority of the households spend a large proportion of their income on food and funeral expenditure. The socioeconomic impacts of HIV/AIDS are far-reaching, especially affecting households relying on a single breadwinner suffering from the infection (Heuveline 2004). For such affected households, a direct consequence is the loss of income source and an increase in health expenditure. Moreover, it burdens older people to look after their grandchildren due to early

deaths of adults in their households. The role of social networks and community-based organizations becomes vital in coping with the socioeconomic problems, which worsens due to prevalence of HIV/AIDS (Foster 2005).

Traditionally, social and kinship networks in rural Sub-Sahara have served as safety nets for communities in coping with a number of stressors together with HIV/AIDS (Pronyk 2002). A primary role of these safety nets is to support affected households in a number of ways. These include mutual support in lending money or food to neighbours, contributing to funerals by the extended family and friends, and providing social support and assistance to friends and members of social clubs. However, numerous studies conducted in the region have hinted that these safety nets or support networks may not remain viable due to increasing number of deaths and household expenditure. Increase in the number of orphans due to high adult mortality has added burden on the extended family and neighbours in taking care of dependents of the dissolved households. Household dissolution may occur if no adult remains in the households to look after the children. A growing number of child-headed households suggest weakening of these safety nets, especially regarding taking care of the orphans.

Unfortunately, majority of the studies addressing the impact of HIV/AIDS on the social fabric (traditional support networks) of rural communities are limited to cross-sectional surveys (Gillespie et al. 2007). As a result, it is not easy to understand the social dynamics related to HIV/AIDS and socioeconomic stressors in the long-term. Moreover, accounts from studies lack precision in describing effect of the epidemic on existing social networks. One way of formally representing this phenomenon is by using agent-based simulation, which is a suitable technique for representing complex social systems. Elsewhere, we have demonstrated how this modelling technique can be useful in describing the socioeconomic impact of HIV/AIDS (Alam 2007a).

In this paper, we suggest a possible way for describing the depletion of social networks due to HIV/AIDS, based on dynamical simulated networks. This suggested approach is based on nonparametric tests for showing changes in networks over time and was described in previous work (Alam 2007b). In subsequent sections, we introduce simulated networks, followed by an overview of the case study for which we have developed an agent-based social simulation model. We will then report and discuss simulation results.

2 Dynamical Simulated Networks and Identifying Structural Changes

Social systems are sources of complexity in themselves in the sense that interactions between individuals can give rise to unexpected and unpredictable behaviour at the system level. One way of understanding the interplay of such interactions is through simulating some aspects of the target system. Agents are the representation of social entities, such as individuals, households and firms. The purpose of agent-based social networks is to explore the simulated data trajectories and the understanding of the modelled phenomena. The methodology used in our work is that the rules of interaction are modelled at the micro level and the resulting network signatures are analyzed at the macro level. This is different as compared to the stochastic models for dynamic social network (c.f. Burk et al. 2007; Snijders et al. 2007), where existing longitudinal data are used for model fitting and parameter estimation (Steglich et al. 2004).

One of the major advantages of agent-based modelling is that it allows incorporating both qualitative and quantitative accounts of dynamic social processes from real case studies (Moss and Edmonds 2005). Evidence-driven agent based modelling constrains agent and mechanism design by independent evidence about the behaviour of relevant actors in the target system. Moreover, this type of modelling facilitates in representing stakeholders', as formal rules for behaviour and social interaction of relevant actors. For instance, terms such as 'vulnerability', 'stressors', 'social fabric' have had been used in different contexts with varying definition. Modelling social networks as 'safety nets' and investigating the effects of deaths attributed to AIDS gives us a formal representation of the notion of vulnerability and stressors.

2.1 Agent-based Models and Generated Social Networks

Agent-based models (ABM) provide a 'methodological' way to capture local interactions among agents who represent social entities. Agents' interactions lead to the formation of ties with other agents and thus the generation of simulated networks. Often, social processes governing the agents' interactions influence the evolution of such networks. ABM is one suitable technique for modelling the interplay of social processes (Carley 2003; Moss and Edmonds 2005), leading to an understanding of emerging network structures. Moreover, where the agents form several types of relations, multiple overlapping networks are generated.

One of the key aspects in our model for exploring impact of HIV/AIDS is that agents' population change due to endogenous birth and death processes. Incorporating rules concerning agents' sexual interactions, marriages, births, progression from HIV to AIDS, and deaths allowed exploring long-term implications of social policies in a rural South African village. Moreover, it allowed modelling both horizontal (via sexual interaction) and vertical (mother-to-child) HIV transmission. Change in agent population, during a simulation run, gives rise to a methodological issue, i.e. comparing networks of different sizes at different time steps. An important implication for comparing dynamical networks is that the network analysis measures used should not depend upon the network size. In the subsequent section, we present a simple technique that is based on comparing distributions of local measures (or agent's characteristics) for an entire population (at a given time step). We use this technique to represent effect of HIV/AIDS in a population at networks generated at individual and households levels.

2.2 Using Kolmogorov-Smirnov Test to Compare Network Snapshots

Agent-based social simulation models are usually analyzed based on a set of hypotheses. Models of descriptive social processes driven from evidence are difficult to analyze. It is even harder in cases where the agents' population is dependent upon the social processes and no longer remains an invariant. Observing the distributions generated from multiple runs of the simulation may help in guessing the system's behaviour in general. It would therefore be imperative to understand the factors that lead to the emergence of networks during a simulation. If one was lucky, one could then identify measures and independent variables that remain valid in most runs.

Nonparametric statistical techniques, and specifically in the context of this paper the Kolmogorov-Smirnov test, are potentially suited in helping to analyze dynamic agent-based networks. These techniques do not assume any prior distribution of the generated data. The classical statistical tests inherently assume that the data comes from the normal distribution. However, the condition for normality does not hold in a number of agent-based models of social processes (c.f. Moss and Edmonds 2005). We present our scheme that is applied to the social networks in this paper. This scheme has been introduced previously in Alam et al. (2007b).

Given a simulation run, we compare network snapshots P= {P0, P0+Δ, P0+2Δ…} where each element Pi represents a distinct population of agents at a particular time in the run, P0 being the population at t=0. For each consecutive

pair (Pi, Pi+Δ), a Kolmogorov-Smirnov (KS) test (Neave and McConwa 1987) is performed. A two-population KS test indicates the likelihood that two datasets come from the same distribution. The p-score (Y-axis) that can be calculated from the KS test can be roughly interpreted as the probability that the two data sets have the same distribution. In our context, a value 1.0 (maximum) implies that there is no change in the degree distribution of the two consecutive snapshots of the agents' network. The procedure as presented by Alam et al. (2007b) is as follows:

- Choose a lag of size Δ for the comparison of network snapshots at different time steps.
- For all time lags ti (i← 0,Δ, 2Δ,…), obtain a corresponding series of the node degrees, Pi.
- For all consecutive pairs (Pi, Pi+Δ) (i← 0,Δ, 2Δ,…), compute the p-score for the KS test for 2-populations.

In this paper, we have considered different snapshots of the same simulation run as separate populations and compared them using the KS test. Our main motivation has been to look for techniques that can help keeping track of the network structure over the course of a simulation run. Observing the p-score for consecutive snapshots can indicate whether the network structure has changed or not. A high p-score is an indicator that there is no evidence from the distributions that the network has changed – a difference is possible but this is unlikely. Moreover, significant differences in the p-score could be one possible indicator in selecting a set of snapshots for further analysis.

3 A Model for Socioeconomic Impact of HIV/AIDS

The scheme introduced in the previous section has been applied to an agent-based model based on a real case study. We model the livelihood and the household structure of a rural community in South Africa. In the case study area, there are socioeconomic stressors to which the local people face, including water scarcity, climate variability, HIV/AIDS, and food insecurity.

The model investigates the impact of HIV/AIDS on households and the overall community structure (Alam et al. 2007a). The case study area is located in the Limpopo region in South Africa and is one of the most vulnerable areas lacking water, food security, jobs and other social infrastructure. Many households have female heads because the men are often living away from the house as migrant workers. State grants are the primary source of income of which a high percentage

is spent on food, health and funeral costs. People try to cope with stressors via mutual help amongst neighbours, friends and extended family. A major concern is the number of orphans in the community that has increased mostly due to HIV/AIDS related adult deaths. The extended family largely accommodates dependents (orphans, old relatives) of a dissolved household (Ziervogel et al. 2006).

The focus of the model is on the behaviour of individual agents as well as that of households and thus attempts to take into account both the individual interactions and the decisions taken by the households. We have adopted a multi-layer network approach to model the social networks. Individuals are represented as agents with a network of friends. Each individual is member of a household, with one of the household members acting as the household head. Households have a network of social neighbours with whom they interact.

Agents and households are created based on the available demographic data. Both endogenous and exogenous factors influence the dynamics of agent interactions. As a result, the size of the generated networks changes over time. Agents are assigned some random friends. With the agents joining the clubs, the size of the friendship network remains dynamic. A high prevalence of HIV/AIDS affects the health of those who are infected resulting in increasing deaths. Orthogonally to the network layers of friends and the household social neighbours is the extended family structure. As stated above, each individual agent is a member of a household. Households in turn form clusters, which represent the extended family. This comes into play when a household dissolves due to the death of all care-providing adults, leaving the dependants (orphan children and possibly any seniors without income) behind. If this happens, an accommodating household has to be found. The dependents search the family hierarchy to determine the nearest living relative who is able to accommodate the surviving dependants. If there is none in the extended family, the search is expanded to the networks of neighbours and friends[1].

4 Simulation Results

The model takes into account several types of networks such as those determined by households' social neighbourhood, extended family ties and agents' friends. There are also savings clubs that are, in fact, fragments of networks and less persistent. The resulting social networks are dynamic and changing over time as well as constrained by the underlying social processes. As agents die and new house-

1 Alam et al. (2007a) present the model and the implemented processes in detail.

holds are created and dissolved, the structure of these networks change as well during the simulation. In this paper, we have applied the scheme discussed in Section 3, only to the degree distribution of the nodes (individual agents and households) in a given network.

In order to investigate changes in the generated social networks, we have applied the technique at the households' social neighbourhood and the agents' friendship network. A high prevalence of the HIV/AIDS in the community implies that agents die at a much earlier age than otherwise and this threatens the viability of the village community in the model.

In our model, creation of a household depends upon marriages of couples and the money required building a new house. It is customary in Sub-Saharan Africa that the bridegroom must pay the bride money (*lobola*) to the bride's household before marriage (Mturi et al. 2003). An important source for arranging this money and contributing towards household's expenditure is the remittances sent by migrant agents in the model. However, as migrant agents are at risk of contracting HIV while away, the number of adult death increases as the simulation proceeds. Death of an adult male agent not only implies that the household has lost a vital source of income, but also, reduces the possibilities of new households being created. Dissolution of households occur when there is no adult member left to look after the children, and this affects the household neighbourhood and kinship network in the model. In other work (Alam et al. 2007a; 2007b), we have shown how processes operating at the micro-level influence the overall network structure (in our case, the household network), and on the other hand, networks constrain agents' decision-making as well.

Different lag sizes have been chosen for comparing the degree distribution of the nodes at consecutive time steps. In this paper, lag sizes of 5, 10, 25 and 50 time steps have been chosen. A single time step corresponds to one month in our model. Simulations were run for 1000 time steps (~ 80 simulation years) and results are reported for 10 runs for each case. The methods implemented for calculating the p-value for the two-population Kolmogorov-Smirnov test have been derived from Press et al. (1992).

In the simulation runs illustrated in Figure 1 and Figure 2, we considered households' social links mapped into a social space. By social space, we mean that the households' neighbourhood defined by their social ties and not by a physical space (e.g. von Neumann neighbourhood). Once the simulation starts, creation and dissolution of the households depend upon by the social processes and assume no a priori distribution.

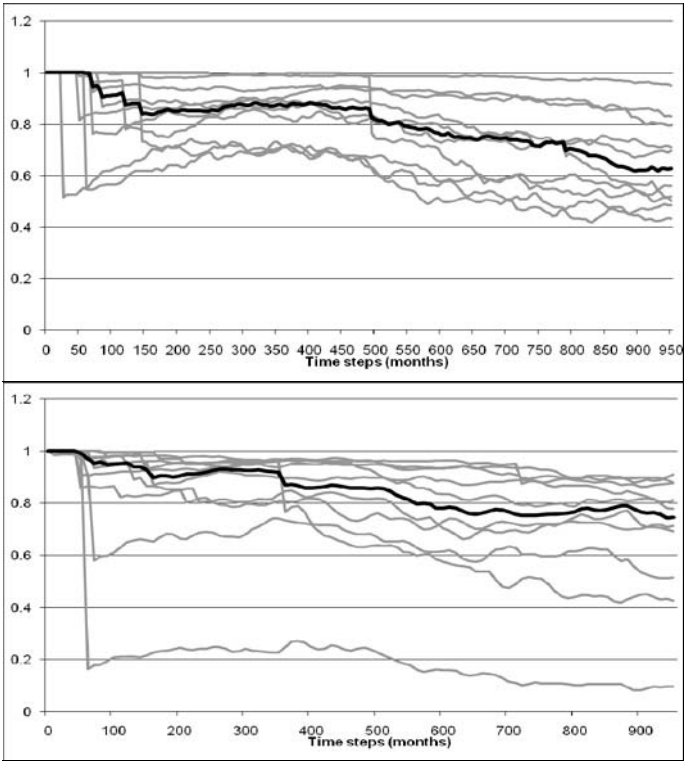


Figure 1: P-score of KS test

Households' social links network snapshot at every (above) 5th and (below) 10th time step respectively, and compared with the previously taken snapshot; black line shows the median of the simulation runs.

In Figure 1 (above) the time lag size was the minimum of the four cases, i.e. 5 time steps. One may observe stable network structure when the networks in this case than in the Figure 1 (below) where the lag size was set as 10. The black dotted line shows the median of the 10 simulation run in each case. The lag size in Figure 2 (above) was set at 25 time steps, while in Figure 2 (below) it was set at 50 time steps. In the two cases presented in Figure 1, the network structures remain stable in most simulation runs. On the other hand, there is a relatively higher variability in the p-score among the 10 simulation runs for Figure 2 (above). Fig-

ure 2 (below), where the time lags were the greatest among the four cases, chances that a pair of network snapshots is similar drops down considerably.

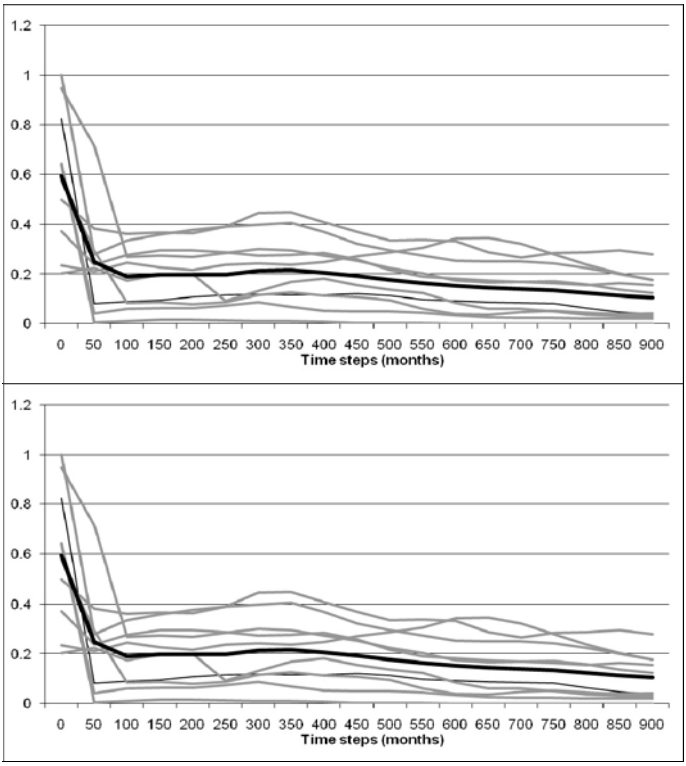


Figure 2: P-score of KS test

Households' social links network snapshot at every (above) 25th and (below) 50th time step respectively, and compared with the previously taken snapshot; the black dotted line shows the median of the simulation runs.

As mentioned above, the social neighbourhood in the model changes when either a new household is created following a marriage, or when a household is dissolved and the occupants being accommodated. Due to an increase in the HIV/AIDS prevalence, agents die earlier but the households continue to exist for a much longer time. A drop in the possibility that two consecutive snapshots of the network are similar for the case of 50 time steps is due to marriages and the creation

of new households. Newly married couples are expected to build new houses at a later stage as their decision to build a new house is constrained by their socioeconomic condition.

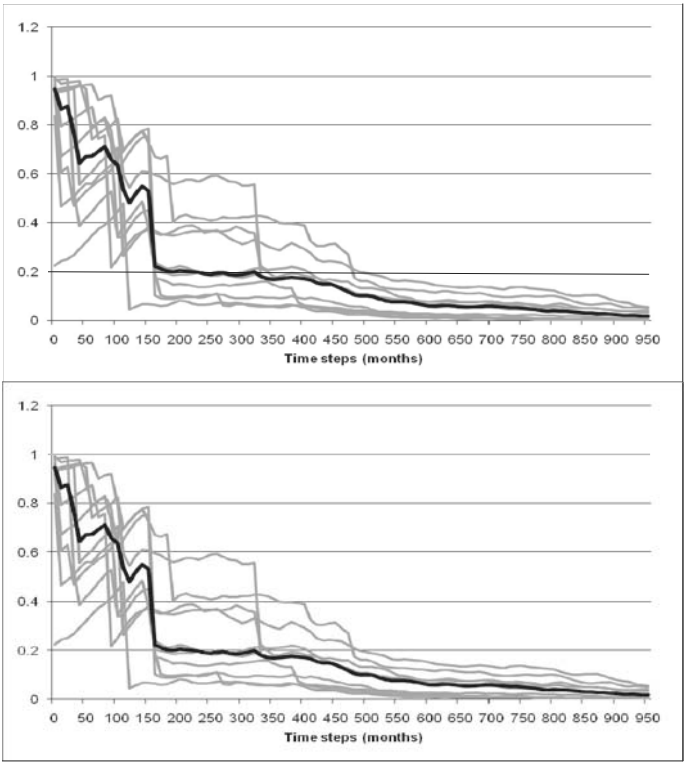


Figure 3: P-score of KS test

Agent's friendship network snapshot at every (above) 5th and (below) 10th time step respectively, and compared with the previously taken snapshot; the black dotted line shows the median of the simulation runs.

Figure 3 illustrates the application of KS test on consecutive snapshots for the agents' friendship network described before. Unlike the households' social neighbourhood, the friendship network changes more rapidly during the simulation run. This is because of the early deaths and consequently agents tend to lose their friends occasionally. Another factor that influences this rapid change is that

agents in our model become friends as they join the savings clubs (discussed before). For larger lags sizes, i.e. 25 and 50th time step, the p-score remained zero for all the comparisons and is therefore not reported here. In sum, the friendship network at the individual level was found to be far less stable than the social neighbourhood household level.

The high variability in the network at the individual level may correspond to anecdotal accounts whereby individuals have lost friends in a short time span or disappearance of individuals from social gatherings such as regular savings clubs or church meetings.

5 Discussion and Outlook

Social systems do not remain in a stable state and are dynamic in nature. Events changing the structure of the network may occur any time during the simulation, which might be missed when using global measures. Agent-based models are validated qualitatively at micro level and the simulated trajectories are analyzed quantitatively. Like those working with real data (Faust 2006), finding suitable measures for comparing networks of different sizes is also a problem for the social simulation community.

In this paper, we have attempted in using the concept of structural changes in a population to explain impact of HIV/AIDS on social networks regulated at individual and household levels. Results reported in the previous section, show greater volatility observed at the agents' friendship network as compared to the social neighbourhood of households in the model. They also suggest that death of individuals in the Sub-Saharan Africa due to HIV/AIDS affects both individuals and households. However, the impact on households and their social and kinship ties may not be appear within a short period as HIV spreads, but may have long-term implications for the sustainability of the community. Our approach provides the basis for future work towards understanding the dynamics related to complex systems using agent-based simulated networks.

For networks of varying size, we have proposed the use of the 2-population Kolmogorov-Smirnov (KS) test in comparing the simulated networks at different time steps. The KS test can indicate when structural changes have occurred in the network during simulation. It takes into account the network as a whole in many different dimensions at once whereas traditional techniques typically provide a *1*-dimensional aggregate measure. This technique is generic and does not restrict to the degree distribution alone. As next steps, we would like to extend this analysis

to other characteristics especially the geodesic distance, which is important to study the role of sexual norms in the spread of HIV/AIDS.

In the social network analysis literature, we find several techniques and model for longitudinal network analysis. Techniques such as Quadratic Assignment Procedure (QAP) have been used in comparing longitudinal networks. QAP is non-parametric and thus requires no a priori assumption about the distribution of the observed data (Krackhardt 1987). However, it requires the networks to be of the same size. Moreover, stochastic dynamic social network models (c.f. Snijders et al. 2007) have so far assumed that the network size remains fixed during two times t_1 and t_2, $t_1 < t_2$. The pairwise t-test is more sophisticated in scrutinizing difference in the population at different time steps; however, the population should remain the same. The nonparametric counterpart for the pairwise t-test is the Wilcoxon signed-rank test; however, that too requires the 'same' population at different time intervals. Recent work by Asur et al. (2007) and Falkwoski et al. (2006) have progressed in analyzing networks where new members join and leave in communities at different times.

References

Alam, S.J., Meyer, R., Ziervogel, G.,& Moss, S. (2007a). The Impact of HIV/AIDS in the Context of Socioeconomic Stressors: An Evidence-driven Approach. *Journal of Artificial Societies and Social Simulation*, 10(4) <http://jasss.soc.surrey.ac.uk/10/4/7.html>.

Alam, S.J., Edmonds, B.,& Meyer, R. (2007b). Identifying Structural Changes in Networks Generated from Agent-based Social Simulation Models. In *Proceedings of Tenth Pacific RIM Intl Workshop on Multi-Agents (PRIMA'07), Bangkok, Thailand.*

Albert, R. & Barabasi, A. (2002). Statistical Mechanics of Complex Networks. *Rev. Mod. Phys.*, 74.

Asur, S., Parthasarathy, S.,& Ucar, D. (2007). An Event-based Framework for Characterizing the Evolutionary Behavior of Interaction Graphs. In *Proceedings of the Thirteenth Conf. on Knowledge Discovery and Data Mining, San Jose, USA*.

Burk, W.J., Steglich, C.E.G.,& Snijders, T.A.B. (2007). Beyond dyadic interdependence Actor-oriented models for co-evolving social networks and individual behaviors. *International Journal of Behavioral Development*, 31, 397-404.

Carley, K.M. (2003). *Dynamic Network Analysis, in Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*. Breiger et al. (Eds.) Washington, DC. 133-145.

Falkowski, T., Bartelheimer, J.,& Spiliopoulou, M. (2006). Mining and Visualizing the Evolution of Subgroups in Social Networks. In *Proceedings of the International Conference on Web Intelligence (WI'06)*. IEEE Press. 52-58.

Foster, G. (2005). *Under the Radar—Community Safety Nets for Children Affected by HIV/AIDS in Poor Households in Sub-Saharan Africa*. United Nations Research Institute for Social Development (UNRISD).

Faust, K. (2006). Comparing Social Networks: Size, Density, and Local Structure. *Metodološki zveski*, 3(2), 185-216.

Gillespie, S., Kadiyala, S.,& Greener, R. (2007). Is poverty or wealth driving HIV transmission? *AIDS*. 21 Suppl 7, 5-16.

Heuveline, P. (2004). Impact of the HIV epidemic on population and household structure: the dynamics and evidence to date. AIDS. 18 (suppl. 2). 45-53.

Kirman, T.W. (1998). *Statistics to Use,* http://www.physics.csbsju.edu/stats/.

Kossinets, G., & Watts, D.J. (2006). Empirical analysis of an evolving social network. *Science*, 311, 88-90.

Krackhardt, D. (1987). QAP Partialling as a Test of Spuriousness. *Social Networks*. 9. 171-186

Moss, S., & Edmonds, B. (2005). Sociology and Simulation: Statistical and Qualitative Cross-Validation. *American Journal of Sociology*, 110(4), 1095-1131.

Mturi, A, Xaba, T.,& Sekokotla, D. (2003). *Assessment of circumstances facing contemporary families in South Africa.* Durban: School of Development Studies, University of Natal.

Neave, H., & McConwa, K. (1987). *Distribution free methods*. Open University Press.

Press, W.H., Flannery, B.P., Teukolsky, S.A.,& Vetterling, W.T. (1992). *Numerical Recipes in C: The Art of Scientific Computing.* 2nd Ed. Cambridge: Cambridge University Press.

Pronyk, P. (2002). *Social Capital and the HIV/AIDS epidemic in rural South Africa: The New Magic Bullet?* The Department of Infectious and Tropical Diseases, The London School of Hygiene & Tropical Medicine, London.
<http://www.wits.ac.za/radar/PDF%20files/socialcapital_HIV.PDF>.

Snijders, T.A.B, Steglich, C.E.G.,& Schweinberger, M. (2007). Modeling the co-evolution of networks and behaviour. In van Montford et al. (Eds.), *Longitudinal models in behavioral and related sciences* (pp. 41-71).

Ziervogel, G., et al. (2006). *Adapting to climate, water and health stresses: insights from Sekhukhune, South Africa.* Stockholm Environment Institute (Oxford).

Contributors

Shah Jamal Alam
Manchester Metropolitan University, Centre for Policy Modelling, United Kingdom, shah.ruth@cfpm.org
Shah Jamal Alam studied computer science at the universities of Karachi and Saarland. He is currently a PhD student at the Centre for Policy Modelling (CPM) supervised by Scott Moss and Bruce Edmonds. His work is supported by the EU FP6 CAVES Project.

Daniel Bochsler
University of Zurich, Center for Comparative and International Studies (CIS), Switzerland, bochsler@ipz.uzh.ch
Daniel Bochsler, is a political scientist at the Universities of Zurich and Geneva. His main research is in quantitative methods and political institutions, and he has been associated with the University of Tartu, University of California at Irvine, and Central European University. He published a monograph on the Swiss cantons: „Die Schweizer Kantone unter der Lupe“ (Haupt, 2004, with co-authors).

John Breslin
National University of Ireland, Digital Enterprise Research Institute (DERI), United Kingdom, john.breslin@deri.org
John Breslin is a researcher at the Digital Enterprise Research Institute (DERI) and an adjunct lecturer with the National University of Ireland, Galway. He researches Semantic Web 2.0 and semantically-interlinked online communities as the leader of the Social Software research group. He has received a number of web awards, including a Golden Spider award for the Irish community site boards which he co-founded in 2000, and two IIA Net Visionary awards in 2005 and 2006.

Thomas N. Friemel
University of Zurich, Institute of Mass Communication and Media Research (IPMZ), Switzerland, th.friemel@ipmz.uzh.ch
Thomas Friemel is an assistant professor at the Institute of Mass Communication and Media Research at the University of Zurich (IPMZ). His interests lie in the

application of social network analysis in communication science with a special focus on interpersonal communication. He served as conference chair of the 3rd and 4th Conference on Applications of Social Network Analysis (ASNA 2006 / 2007) and is the editor of the respective proceedings (UVK 2007; VS Verlag 2008).

Andreas Harth
National University of Ireland, Digital Enterprise Research Institute (DERI), United Kingdom, andreas.harth@deri.org
Andreas Harth is a PhD student with the Digital Enterprise Research Institute (DERI) at the National University of Ireland, Galway. He holds a Dipl.-Inf. (FH) from Fachhochschule Würzburg, Germany. Andreas worked at Fraunhofer Gesellschaft in Würzburg, at IBM in San Jose, CA, and at USC's Information Sciences Institute in Marine del Rey, CA. His research interests are large-scale data interoperation on the Web, peer-to-peer systems, knowledge representation, and computational logic.

Conor Hayes
National University of Ireland, Digital Enterprise Research Institute (DERI), United Kingdom, conor.hayes@deri.org
Conor Hayes is a senior post-doctoral research fellow at the Digital Enterprise Research Institute (DERI) and an adjunct lecturer with the National University of Ireland, Galway. He has worked as a post-doctoral researcher at ITC-IRST, Trento, Italy and Trinity College Dublin. His research interests include recommender systems, data-mining, information retrieval and modelling of information flows in blogs and bulletin boards.

Debra Hevenstone
University of Michigan, USA & ETH Zurich, Switzerland, dhevenst@umich.edu
Debra Hevenstone is a PhD candidate in sociology, public policy, and complex systems at the University of Michigan visiting the sociology department at ETH in Zurich, Switzerland. Her current research focuses on atypical employment relationships. She has also worked as a public policy analyst at the Brookings Institution.

John Judge
IBM Language Ware, Dublin Software Lab, Ireland, johnjudge@ie.ibm.com
John Judge, PhD, is a researcher at the IBM Dublin Software Lab, working as part of the LanguageWare team. His current duties include research and development

for the NEPOMUK project. Dr. Judge is the joint author of six peer-reviewed research publications and has one patent pending. His research interests include natural language processing and socio-semantic web applications.

Sheila Kinsella
National University of Ireland, Digital Enterprise Research Institute (DERI), United Kingdom, sheila.kinsella@deri.org
Sheila Kinsella is a PhD student with the Digital Enterprise Research Institute (DERI) at the National University of Ireland, Galway. She holds a B.E. in Electronic and Computer Engineering from the National University of Ireland, Galway. Her research interests include the Semantic Web and analysis of object-centered social networks.

Victor V. Kryssanov
Ritsumeikan University, Faculty of Information Science and Engineering, Japan, kvvictor@is.ritsumei.ac.jp
Victor V. Kryssanov is associate professor at the Faculty of Information Science and Engineering, Ritsumeikan University, Japan. He is the author of a great number of scientific and technical papers, and his current research interests include modelling complex systems and the application of physics methods to social network analysis.

Evgeny L. Kuleshov
Far-Eastern National University, Institute of Physics and Information Technologies, Russia, kuleshov@lemoi.phys.dvgu.ru
Evgeny L. Kuleshov is professor and Chair of Computer Systems at the Far-Eastern National University, Russia. He is a distinguished expert in the fields of stochastic process modelling and statistical analysis.

Ruth Meyer
Manchester Metropolitan University, Centre for Policy Modelling, United Kingdom, shah.ruth@cfpm.org
Ruth Meyer graduated in computer science and biology from Hamburg University, Germany, and is in the final stages of completing her PhD in agent-based simulation. She is currently a research associate at the Centre for Policy Modelling, working for the CAVES project.

Gerald Mollenhorst
Utrecht University, Department of Sociology, The Netherlands, g.w.mollenhorst@uu.nl
Gerald Mollenhorst is a PhD candidate in sociology at Utrecht University and the Interuniversity Center for Social Science Theory and Methodology (ICS). His research focuses on the social contexts in which people meet and engage in personal relationships.

Georg P. Mueller
University of Fribourg, Department of Social Science, Switzerland, georg.mueller@unifr.ch
Georg P. Mueller is Maître d'Enseignement et de Recherche (senior lecturer) at the University of Fribourg, Switzerland. His scientific interests include the methodology of social research, with a special focus on the formalization of qualitative analyses; the construction and use of social indicators; and the simulation and mathematical modelling of social processes on the grounds of rational choice and game theory.

Hitoshi Ogawa
Ritsumeikan University, Faculty of Information Science and Engineering, Japan, ogawa@is.ritsumei.ac.jp
Hitoshi Ogawa is professor at the Faculty of Information Science and Engineering, Ritsumeikan University, Japan. His major fields of specialization include multi-agent system theory and applications, and constraint-satisfaction-based resolution methods.

Frank J. Rinaldo
Ritsumeikan University, Faculty of Information Science and Engineering, Japan, rinaldo@is.ritsumei.ac.jp
Frank J. Rinaldo is professor at the Faculty of Information Science and Engineering, Ritsumeikan University, Japan. His major areas of research are AI applications and computer games.

Wolfgang Sodeur
University Duisburg-Essen, Germany, wolfgang.sodeur@t-online.de
Wolfgang Sodeur holds a PhD in sociology (University of Cologne, 1970) and taught at the universities of Hamburg (1972/73), Wuppertal (1973-1988), and Essen (1987-2004) as a professor for empirical social research. He published books

about small group research and numerical classification. Current research interests include network research, socialization processes, and regional statistics.

Mikhail Sogrin
IBM Language Ware, Dublin Software Lab, United Kingdom,
sogrimik@ie.ibm.com
Mikhail Sogrin is a researcher and software engineer at the LanguageWare team in IBM Dublin Software Lab. He participated in, and took fifteenth place in, the World Finals of 2000 ACM International Collegiate Programming Contest. His current duties include research and development for the NEPOMUK project. Mr. Sogrin is a joint author of five peer-reviewed research publications and has one patent pending.

Volker G. Täube
European Commission, Eurostat, Luxembourg, volker.taeube@ec.europa.eu
Volker G. Täube holds a PhD in sociology (University of Essen, 2001) and is working as a Seconded National Expert on behalf of the European Free Trade Association (EFTA) for the European Commission (Eurostat). His current scientific interests are structural analysis, social capital measurement, social network analysis, and information technologies.

Alexander Troussov
IBM Language Ware, Dublin Software Lab, United Kingdom,
atrousso@ie.ibm.com
Alexander Troussov, PhD, is chief scientist at the IBM Dublin Centre for Advanced Studies (CAS) and chief scientist of the IBM LanguageWare group. He is a joint author of more than 30 peer-reviewed research publications and has five patents pending. As CAS chief scientist, he leads IBM's participation in the NEPOMUK project.