Methods for Collection of Participant-aided Sociograms for the Study of Social, Sexual and Substance-using Networks Among Young Men Who Have Sex with Men

Lisa M. Kuhns
Ann & Robert H. Lurie Children's Hospital of Chicago & Northwestern University Chicago, IL USA

M. Birkett
Northwestern University Chicago, IL USA

B. Mustanski
Northwestern University Chicago, IL USA

S.Q. Muth
Quintus-ential Solutions Colorado Springs, CO USA

C. Latkin
Johns Hopkins Bloomberg School of Public Health Baltimore, MD USA

I. Ortiz-Estes
Ann & Robert H. Lurie Children's Hospital of Chicago Chicago, IL USA

R. Garofalo
Ann & Robert H. Lurie Children's Hospital of Chicago & Northwestern University Chicago, IL USA

Abstract
In this study, we adapted and tested a participant-aided sociogram approach for the study of the social, sexual, and substance use networks of young men who have sex with men (YMSM); a population of increasing and disproportionate risk of HIV infection. We used a combination of two interviewer-administered procedures: completion of a pre-numbered list form to enumerate alters and to capture alter attributes; and a participant-aided sociogram to capture respondent report of interactions between alters on an erasable whiteboard. We followed the collection of alter interactions via the sociogram with a traditional matrix-based tie elicitation approach for a sub-sample of respondents for comparison purposes. Digital photographs of each network drawn on the whiteboard serve as the raw data for entry into a database in which group interactions are stored. Visual feedback of the network was created at the point of data entry, using NetDraw network visualization software for comparison to the network structure elicited via the sociogram. In a sample of 175 YMSM, we found this approach to be feasible and reliable, with high rates of participation among those eligible for the study and substantial agreement between the participant-aided sociogram in comparison to a traditional matrix-based approach. We believe that key strengths of this approach are the engagement and maintenance of participant attention and reduction of participant burden for alter tie elicitation. A key weakness is the challenge of entry of interview-based list form and sociogram data into the database. Our experience suggests that this approach to data collection is feasible and particularly appropriate for an adolescent and young adult population. This builds on and advances visualization-based approaches to social network data collection.

Keywords: *Sociogram, MSM, youth, egocentric network*

Authors
Lisa M. Kuhns, Ann & Robert H. Lurie Children's Hospital of Chicago, Division of Adolescent Medicine, Chicago, IL. USA and Northwestern University, Feinberg School of Medicine, Department of Pediatrics, Chicago, IL USA.

M. Birkett, Northwestern University, Feinberg School of Medicine, Department of Medical Social Sciences, Chicago, IL USA.

DOI: http://dx.doi.org/10.17266/35.1.1

S.Q. Muth, Quintus-ential Solutions, Colorado Springs, CO USA.

C. Latkin, Johns Hopkins Bloomberg School of Public Health, Department of Health, Behavior and Society, Baltimore, MD USA.

I. Ortiz-Estes, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL USA.

R. Garofalo, Ann & Robert H. Lurie Children's Hospital of Chicago & Northwestern University, Chicago, IL USA.

Acknowledgements
We thank Sarah Brewster and Katie Andrews for their assistance with data collection and management and Crew 450 study participants for their time and effort. The project described herein was supported by grants from the National Institute on Drug Abuse: R01DA025548 and R01DA025548-S (PIs: R. Garofalo, B. Mustanski) The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Drug Abuse or the National Institutes of Health.

Correspondence concerning this work should be addressed to Lisa M. Kuhns, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, 225 E. Chicago Ave. #161, Chicago, IL 60611.

1. Introduction

In its third decade, the HIV epidemic continues to disproportionately affect men who have sex with men (MSM), but has shifted to increasingly affect young MSM (YMSM). In the United States (US), approximately 49,000 new HIV infections occur each year, with a third of incident cases occurring in youth below the age of 30 (Centers for Disease Control and Prevention, 2013a). Between 2008-2010, male-to-male sexual contact accounted for over half of estimated new infections annually; young Black MSM aged 13–24 accounting for more new infections than any other age group or race of MSM (Centers for Disease Control and Prevention, 2013b). Despite ongoing educational efforts, YMSM continue to engage in high risk sexual behaviors placing them at-risk for HIV (Mustanski, Newcomb, Du Bois, Garcia, & Grov, 2011).

Thus, over the past two decades a shift has occurred in the study of the epidemiology of sexually transmitted infections from a focus primarily on individual risk factors to a focus which includes characteristics of social and sexual networks (Aral, 1999; Wohlfeiler & Potterat, 2005). It has been proposed that social networks form part of a web of health causation, with social-structural conditions affecting the formation of social networks, which in turn transfer their influence to health through several basic pathways: social support, social influence, access to resources, social involvement, and person-to-person contact/contagion (Berkman & Glass, 2000; Smith & Christakis, 2008). These psychosocial and behavioral processes then impact health through more proximate mechanisms, including psychological stress responses and health behaviors (Berkman & Glass, 2000). In addition, sexual network ties confer risk for STI/HIV infection through member characteristics and behaviors (Aral, 1999). Among adolescents and young adults, partner characteristics such as age discordance, previous incarceration, STI diagnosis in the past year, other partners in the past year, and problems with alcohol and drugs have been found to influence individual STI/HIV risk (Mustanski, Newcomb, & Clerkin, 2011; Newcomb & Mustanski, 2013; Staras, Cook, & Clark, 2009). Mathematical modeling studies have found that the structural characteristics of sexual networks, including patterns of mixing (Anderson, Gupta, & Ng, 1990), concurrency (Morris & Kretzschmar, 1997), and degree (Christley et al., 2005) impact disease spread, and that position within the network influences infection/transmission (Christley et al., 2005). While networks in which transmission takes place have a common network structure, the actual level of transmission may be best determined by studying factors specific to a population group (Rothenberg & Muth, 2007).

Given the epidemic of HIV among YMSM, social network approaches to the study of HIV risk are an important area of focus (Clatts, Goldsamt, Neaigus, & Welle, 2003), although methods for network data collection developed and tested among YMSM are limited. YMSM constitute an unbounded, hidden, and stigmatized population which present challenges for the collection of network data. While methods of network-based recruitment, such as respondent-driven sampling (RDS; Heckathorn, 1997, 2002) have been developed

for stigmatized and hidden populations and have been implemented with some success to sample MSM (see for example: Iguchi et al., 2009; Ramirez-Valles, Heckathorn, Vazquez, Diaz, & Campbell, 2005; Reisner et al., 2010; Rhodes et al., 2012; Schneider, Michaels, & Bouris, 2012), RDS was developed specifically for sampling hidden populations, not for the elucidation of the immediate network environment per se (see Dennis et al., 2013 as an exception). Given the unbounded nature of the population, the few studies of YMSM networks that have been completed have used egocentric network approaches (Clerkin, Newcomb, & Mustanski, 2011; Kapadia et al., 2013). This type of network study is called an "egocentric" study because all network information is derived from respondent, or "ego" perceptions (Marsden, 1990), rather than from firsthand reports from all individuals in the network.

In this paper, we describe the development and testing of a novel participant-aided sociogram approach for a study of the personal networks of YMSM, assess the feasibility of data collection using this approach, and describe personal network characteristics of the target population. We define social networks as a set of nodes (e.g., persons) linked by a set of social relationships (e.g., friendships) of a specified type (Laumann, Galaskiewicz, & Marsden, 1978). A sociogram is a diagram used to represent the relationships between individuals in a group (Moreno, 1953). The sociogram creates a "picture" in which persons are represented as points and relationships are represented by lines between these points.

The sociogram, or network diagram, has become a mainstay of social network visualization (Wasserman & Faust, 1994); however, as a key tool of social network analysis, the network diagram has primarily been created after data collection has been completed. This is due in part to the focus in much of network analysis on whole networks, i.e., mapping networks of connections among members of entire populations (Hogan, Carrasco, & Wellman, 2007). More recently, methods have developed for bringing the sociogram into the data collection process for collection of personal egocentric network data, i.e., in which the respondent (or "ego") provides their direct contacts (or "alters") as well as relationships between those alters (Hogan et al., 2007; Kennedy, Tucker, Green, Golinelli, & Ewing, 2012; Tucker et al., 2012). This process has advantages, including a high level of participant engagement in the network elicitation process, the immediate visual feedback, and verification of network structure by the respondent and the acceleration of the process of identifying ties between alters (Hogan et al., 2007).

Traditional approaches to elicit network connections require that the ego identify connections between each pair of alters, which is both labor-intensive for interviewers and tedious for respondents. Hogan and colleagues (2007) for example, developed a method to structure the network data as it is being collected in the form of real-time visualization. In their study of communication media and its impact on personal networks in Canada, a detailed network interview was completed with adults in which network members (i.e., somewhat to very close ties) were generated, names transcribed onto "Post-it" notes, and then arranged on a large sheet of paper. Alter ties were elicited by drawing circles around alters rather than completing all possible pairs. They found that this low-technology and interactive method improved interview quality, and was more time and cost-efficient than traditional matrix-based approaches that seek to capture all alter pairs (Hogan et al., 2007).

Participant-aided visualization techniques have also been used among adolescent and young adult populations, specifically homeless youth at risk of HIV infection, in which network characteristics and structure are particularly salient (Rice, Barman-Achikari, Milburn, & Monro, 2012). Among homeless youth seeking services at shelters, drop-in centers, and street venues, Kennedy and colleagues (2012) completed network interviews in which names of social network contacts (i.e., "individual they knew, who knew them, and with whom they had contact during the past year or so") were generated, alter characteristics were elicited, and then a traditional pair-based approach was used to elicit alter ties. Sexual partners were identified from among network members and questions about sexual risk were elicited about each partner. Visualizations of personal networks were then produced immediately using visualization software, and participants were asked to describe distinct clusters of alters (components) as well as alters with no connections (isolates) based on these visualizations. Composite indicators of network structure developed from these techniques were then included in multi-level models to analyze their impact on sexual risk (Kennedy et al., 2012). Tucker and colleagues used a similar name generator in a study of homeless YMSM more specifically, i.e., "name people that you know and who know you and that you had contact with in the past 3 months" to generate a list of network members and elicited the perceived risk behaviors of alters (including sexual risk behavior and alcohol and illicit drug use), however no alter-to-alter ties were elicited, therefore analyses of network structure were limited (Tucker et al., 2012).

Thus, while participant-aided visualization techniques have been developed for health-related

research and utilized with young populations to measure network structure, some evidence suggests that they type of visualization method used is important. For example, McCarthy and colleagues (2007), among a small sample of adults, compared a freestyle network drawing technique such as that used in the Hogan study (2007) to a visualization approach based on a matrix-based elicitation of ties. Participants were instructed to first name forty-five alters with the instructions: "you know them and they know you, by sight or by name. You have had some contact with them in the past two years – by phone face-to-face, email, mail – and you could contact them again if necessary." Respondents were instructed to create a representation of their network by drawing "social circles" and labeling them to describe each group. Respondents then returned no later than a week later to complete a matrix of pairwise ties, which was mapped using visualization software. Study investigators found that the freestyle drawing often produced fewer and/or more homogenous ties than the visualization based on pairwise ties (McCarty, Molina, Aguilar, & Rota, 2007). The investigators concluded that cognitive information on social network ties is not stored randomly, but arranged in patterns that aid recall and which correspond to social structure. Because this "shorthand" method of visualization may result in a consolidation of ties and thus a potential loss of data, these results call for further comparison of the two approaches in other populations and within larger samples.

2. Methods

2.1 Participants

In the study described herein, we measured network structure in the context of an ongoing longitudinal study of a syndemic of health issues facing YMSM in Chicago: the nation's third largest city and the epicenter of the HIV/AIDS epidemic in the Midwest. The purpose of the parent study is to characterize the prevalence, course, and predictors of a syndemic of health problems among YMSM. This syndemic includes substance abuse, experiences of violence, sexual risk taking, and internalizing mental health problems, which increases risk for HIV/STIs. We used a modified form of RDS to enroll YMSM between ages 16 and 20 in the parent study (Kuhns et al., 2014).

Participants for the network sub-study were recruited from the parent study at either the 12- or 24-month follow-up visits (i.e., at the 24-month visit if the 12-month visit had already been completed during the period of sub-study) from June 2011-October 2012. We chose these time points for strategic reasons: 1) to allow for sufficient prior interaction/trust-building between the respondent and the research team to facilitate collection of sensitive network-based information and 2) to coincide with HIV and STI testing, which is completed at baseline and 12-month intervals thereafter. The target sample size for the study (N=175) was determined based on an a priori power analysis linked to analysis goals (i.e., for multilevel analyses, see Birkett, Kuhns, Latkin, Muth, & Mustanski, in press; Mustanski, Birkett, Kuhns, Latkin, & Muth, 2014).

2.2 Procedures

Because studying whole networks of YMSM would be impractical given limited resources (Latkin, Forman, Knowlton, & Sherman, 2003; Potterat, Woodhouse, Muth, & et al., 2004) we chose an egocentric or personal network data collection approach, gathering secondhand information about the immediate network environment from the respondents' viewpoint. Similar to prior participant-aided visualization approaches described above, our network data collection procedures included three processes to elicit network-based information. This included asking the respondent to enumerate all persons with whom they have a certain type of connection, to describe characteristics of those individuals, and to describe social, substance use and sexual connections between these individuals. We used a combination of two interviewer-administered procedures to collect these data: 1) completion of a pre-numbered list form (i.e., in paper-and-pencil format) to enumerate alters and to capture alter attributes; and 2) a participant-aided sociogram to capture respondent report of interactions between alters (sexual, substance-using, and social networks).

2.3 Measures

Name Generators. The enumeration of alters using name generators has been used extensively in network-based approaches (Wasserman & Faust, 1994) and has been found to be reliable for reports of individuals with whom the ego has repeated and salient interactions (Freeman, Romney, & Freeman, 1987). Our name generators and alter characteristic elicitation questions and procedures were based on prior studies of populations at risk of HIV infection (Auerswald, Muth, Brown, Padian, & Ellen, 2006; Latkin & Knowlton, 2005; Potterat, Rothenberg, & Muth, 1999; Potterat et al., 2004; Rothenberg, Baldwin, Trotter, & Muth, 2001). Egos were asked to elicit up to 40 alters on a pre-numbered list form, and provide sufficient identifying characteristics for tracking purposes (e.g., full

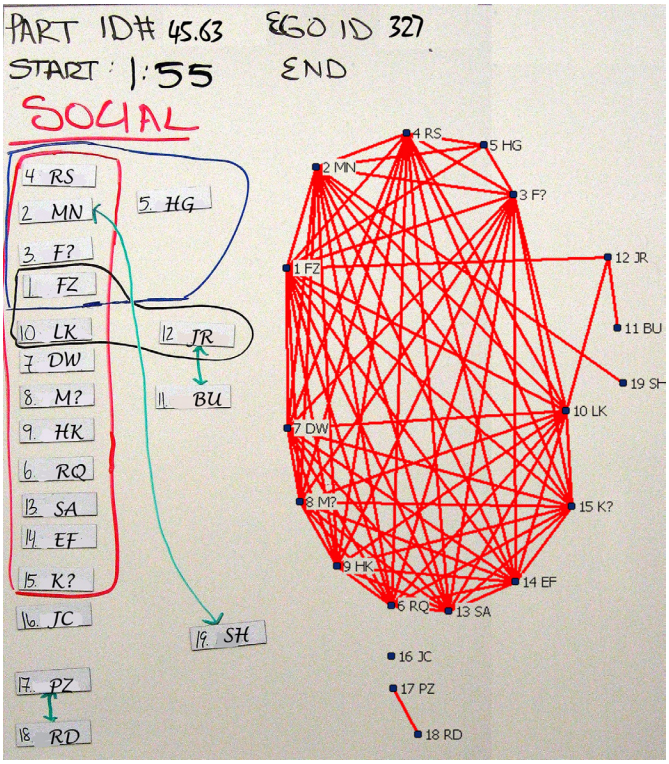
names or failing that, nicknames or initials). We used a set of name generators to elicit social network members who provide emotional and/or instrumental support, including support for sexual minority-related issues:

a. Name the people you are closest to, that is, people you see or talk to regularly and share your personal thoughts and feelings with.
b. Can you think of other people who would give time and energy to help you?
c. Can you think of other people who you could count on to lend or give you $25 or something of equal or greater value?
d. Can you think of other people who you could turn to for help or advice about gay-related issues or problems (for example, if you were being harassed)?
e. Can you think of other people you spend time with on a regular basis yet are not very close to you?

We also asked whether or not the respondent had used substances with or had sex with any additional individuals not listed, or if they knew of other individuals who had used substances or had sex with two or more network members in order to identify additional substance-using and sexual partners not in the respondents' social network (i.e., to elicit more complete substance-using and sexual networks) and to identify individuals in networks that might be in central positions in terms of sex and substance use behavior, but only weakly tied to the ego.

Name Interpreter. We then used a structured name interpreter to elicit demographic and behavioral characteristics of alters. Because signaling to respondents that large amounts of data would be collected on each alter might dampen enthusiasm for providing a complete list of names, we separated the name generator (on page 1 of the list form) from the questions on alter characteristics and relationships (pages 2-4 of the list form). This also facilitated a higher degree of confidentiality; we decoupled alter names from characteristics and securely stored them separately. Alter characteristics and relationships between the ego and alters collected in the name interpreter included: frequency of communication within the last 6 months (0=none to 5=daily), strength of the relationship (1=very close, 2=somewhat close, 3=not at all close), type of relationship (e.g., family member, friend, co-worker), estimated age (in years), race, gender, sexual orientation, and residential location (i.e., nearest cross-streets). We pilot tested use of the list form in multiple modalities and ultimately determined that collecting characteristics alter-wise, rather than question-wise, was more efficient.

Figure 1: Sample social network drawing (left) and as converted to Netdraw figure (right)

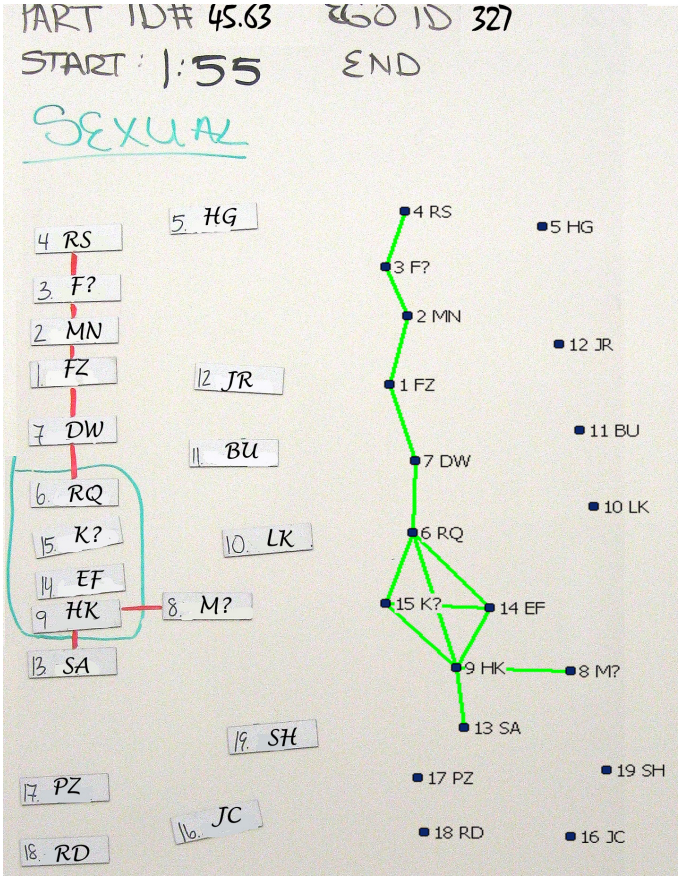


*Participant and ego ID and alter initials were changed to protect participant confidentiality. Colors and arrows in participant drawing were used only to distinguish connections between alters (i.e., to facilitate data entry, not to depict different types of connections or directionality).

In addition to demographic items, we included an additional set of questions regarding sexual and substance-using behavior between the ego and each alter, including: whether or not drugs or alcohol were ever used, frequency of drug or alcohol use in the last 6 months (0=never, 1=1-2 times, 2=once/month or less, 3=2-3 times/month, 4=1-2 times/week, 5=3-5 times/week, 6=every day/almost every day), substances used in the last 6 months (list of 17 substances including alcohol), sexual contact at any time in the past, date of first/last sex, frequency by type of sex in the last 6 months (i.e., oral, vaginal, anal; 0=never, 1=1-2 times, 2=once/month or less, 3=2-3 times/month, 4=1-2 times/week, 5=3-5 times/week, 6=every day/almost every day) and frequency of condom use in the last 6 months (1=always, 2=more than half the time, 3=about half the time, 4=less than half the time, 5=never).

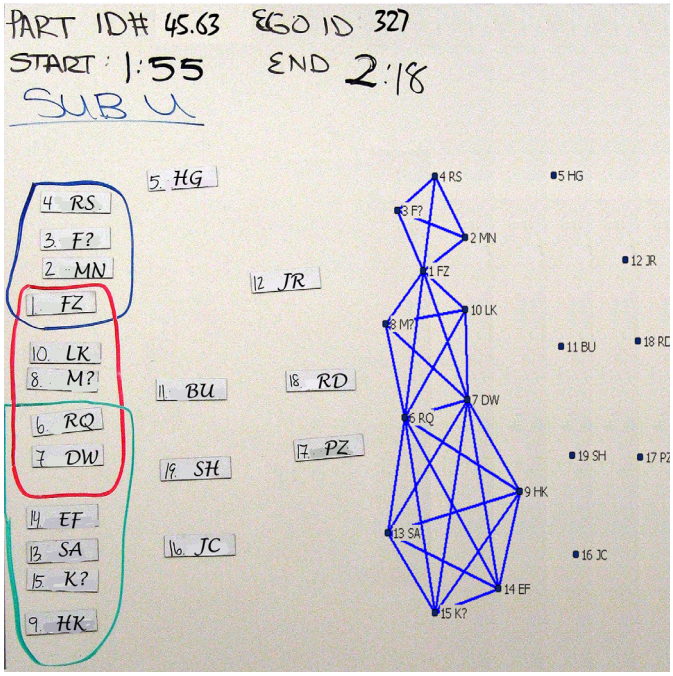
Sociogram. After the name generator and name interpreters were completed, each respondent completed a sociogram to identify social, sexual, and substance-using connections between alters. Whereas Hogan and colleagues used a paper sociogram form with "post-it" notes to indicate alters, we chose a slightly more durable

Figure 2: Sample sexual network drawing (left) and as converted to Netdraw figure (right)*



*Participant and ego ID and alter initials were changed to protect participant confidentiality. Colors in participant drawing were used only to distinguish connections between alters (i.e., to facilitate data entry, not to depict different types of connections).

Figure 3: Sample substance using network drawing (left) and as converted to Netdraw figure (right)*



*Participant and ego ID and alter initials were changed to protect participant confidentiality. Colors in participant drawing were used only to distinguish connections between alters (i.e., to facilitate data entry, not to depict different types of connections or directionality).

approach using erasable, writable magnets, and an erasable whiteboard (see Figures 1-3). We made this modification for our young target population because we anticipated more movement of alters around the whiteboard until the respondent settled on positions, and thus a need for more durable materials (i.e., we anticipated post-it notes losing their stickiness with repeated movement). Alter first names and last initial were transferred to numbered "write-on" magnets (i.e., pre-numbered with the corresponding alter identification number on the list form) and placed these on a magnetic whiteboard. Following Hogan's method, we instructed respondents to arrange alters such that those who know each other were placed together, to draw a circle around any group of three or more alters who know each other well (to identify cliques) and to draw a line between any two alters (outside of a clique) who know each other well (to identify dyads). A digital photograph of this social network sociogram was taken (and verified using the view function), the white board was erased, and this process was completed for each of the other networks of interest – the substance-using and sexual networks. In these cases, respondents were directed to draw circles around respondents and lines between respondents who were known to the respondent to have used substances together or had sexual relations.

Database. Data from the name generator, name interpreter, and sociograms were then entered into a database programmed using Visual Basic in Microsoft Access. Groups of interactions depicted on the sociogram (both cliques and dyads) were directly entered into this database, which was programmed to automatically convert the groups to dyadic lists. The database was formatted such that the data entry tabs matched the name generator and name interpreter forms to facilitate data entry. Once names from the list form were initially entered, the remaining database forms were pre-populated with initials to facilitate data entry. Finally, instantaneous visual feedback of the network was provided at the point of data entry, with custom Visual Basic programming creating and launching NetDraw VNA files (Borgatti, 2002) directly from the database, facilitating verification.

Assessment of reliability. Given prior evidence suggesting more comprehensive data elicited via matrix-based visualization versus the freestyle drawing approach cited above, we sought to test the reliability of a freestyle method of tie elicitation in comparison to the matrix-based elicitation. The freestyle drawing approach reflects, in effect, a "shortcut" to tie elicitation: a respondent might forget or otherwise fail to describe ties between individuals when not specifically asked

about them. From a classic psychometric theory approach to measurement, measures are reliable to the extent that they are repeatable and free from error when persons, instruments, and conditions vary (Nunnally & Bernstein, 1994). In this case, we were interested in the consistency of alter-to-alter tie information when elicited using two different instruments – the participant-aided sociogram method and a traditional matrix-based approach. Thus, in a subsample of 52 participants, we assessed the reliability of reported social ties through the sociogram method in comparison to a matrix-based approach. Following initial data collection, we instructed participants to take a 5-10 minute break; we then elicited ties (between alters) using pairwise comparison from a random sample of 5 alters taken from the respondent's list form. We asked respondents to indicate whether each pair of alters was socially connected, i.e., if they knew each other well, if they had a sexual relationship and if they had used substances together. We compared agreement between the two approaches using a coefficient of agreement, Cohen's kappa (Cohen, 1960), calculated for each respondent with a mean value calculated across all respondents for each type of network.

3. Results

Study enrollment was completed between June of 2011 and October of 2012. The study protocol was approved by the Institutional Review Boards at each site and respondents were provided with $25 for participation. A total of 204 parent study participants were approached for participation, and 179 (88%) agreed to participate, however, two participants did not show up for scheduled appointments and two were enrolled, but subsequently withdrew. The final sample size was 175 (86%). The reasons cited for non-participation included both logistical issues, e.g., no longer living in the area and scheduling conflicts, but also included lack of interest. The two participants who withdrew cited privacy concerns as their reasons for withdrawal and asked to have any network data destroyed. The demographics of enrollees are parallel to those of the parent study in terms of age, race, and sexual orientation. Participants had a mean age of 20.1 (SD=1.4; range= 17.1 to 23.0) and were majority African American (53.7%) and gay-identified (82.9% identified as completely/mostly gay). Interviews lasted approximately 55 minutes on average (SD=21.5 minutes; range=19-129 minutes).

3.1 Name Generator and Alter Distribution

The name generator elicited a total of 2,579 core social network alters within 175 personal networks, with an average of 14.7 alters (SD=8.4; range=3-40) reported per participant (76.1% of all alters named; see Table 1). Asking participants to list any additional network members with whom they had sex or used substances (but had not yet named) resulted in generation of 689 (20.3%) more names in the network. When asked about network members not directly tied to the ego, but who had sex or used substances with at least two network alters resulted in the generation of 122 (3.6%) more names (see Figure 4 for a visualization of these ego networks with tie type differentiated by node size). Only one respondent listed the maximum of 40 alters. The initial name generator, "Name the people you are closest to, that is, people you see or talk to regularly and share your personal thoughts and feelings with," elicited the greatest number of responses, 1,008 alters (39.1%), followed by the second question (i.e., Can you think of other people who would give time and energy to help you?"), which generated 645 alters (25.0%). The name generator specific to gay-related support (i.e., Can you think of other people who you could turn to for help or advice about gay-related issues or problems, for example, if you were being harassed?) resulted in 262 additional alters (10.2%). Respondents identified social network alters largely by both first and last name (93.1%), with the next highest percentage being first name only (5.0%). All participants reported at least a first, last, or nickname for social network alters.

Respondents' social network alters included primarily friends (60.0%), followed by family members (23.8%) and those identified as "acquaintance/associate" (4.7%). The strength of relationships was reported to be 'very close' to 'somewhat close' on average (i.e., mean=1.7; SD=0.7; range 1=very close, 3=not at all close). Respondents communicated with members of their networks (in the last 6 months) an average of weekly to a couple of times a monthly (i.e., mean=3.74; SD=1.3; range 0=none, 5=daily).

3.2 Alter Characteristics

In terms of age, social network alters were largely in the 19-22 year old age group (46.3%), followed by 23-29 (17.9%); very few alters were 18 or younger (14.4%; see Table 2). By race, alters were primarily African American (45.0%), followed by White (23.8%) and Latino (22.1%). Racial homophily across the core networks was 83.0% among African American respondents, 73.0% among Whites and 67.9% among Latinos. Respondents named alters of both sexes (50.7% male) and in terms of sexual orientation, gay (36.8%) or heterosexual (49.7%) alters. Residential location data was provided for the vast

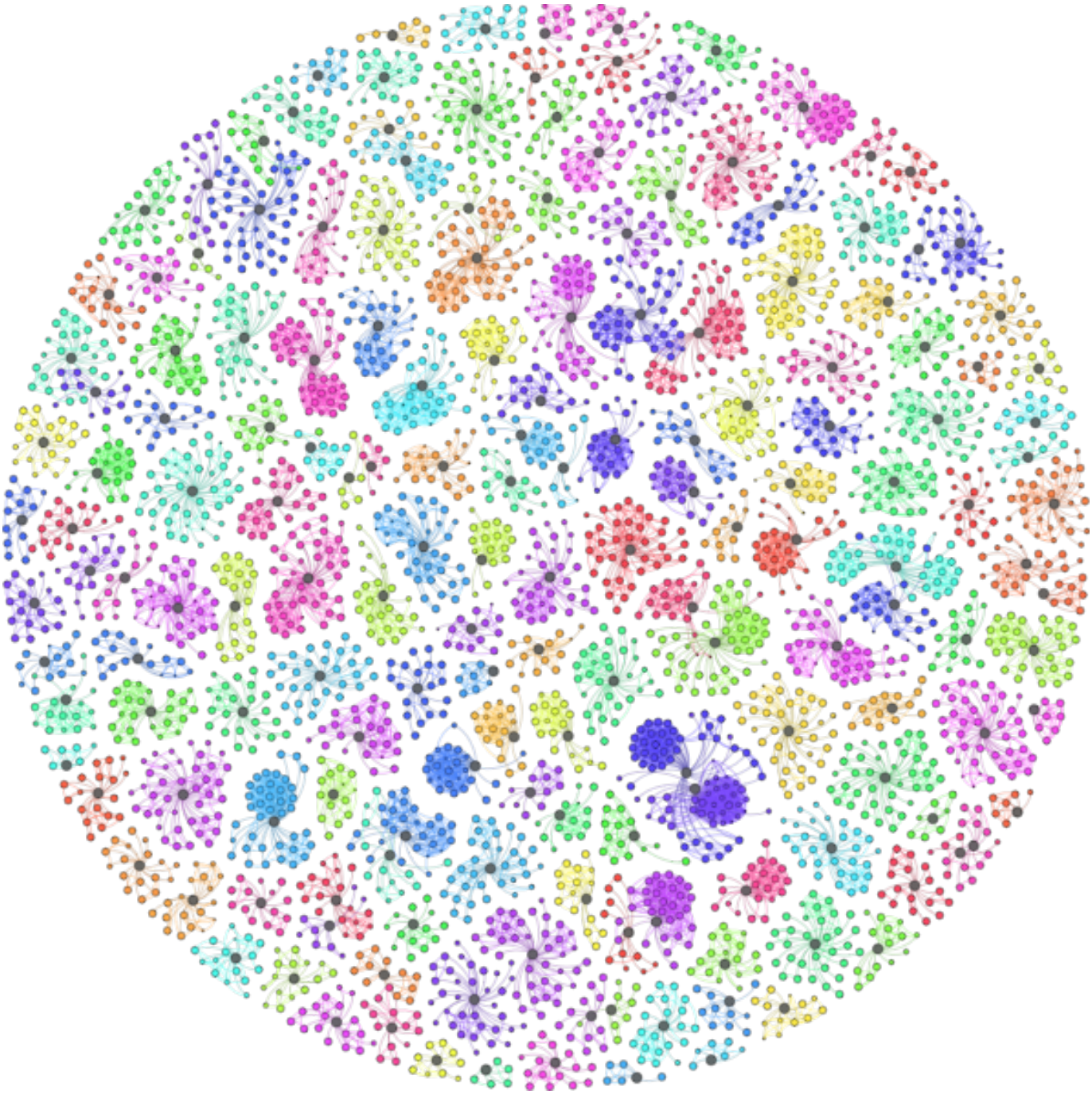


Figure 4: Ego = dark gray color, node size 4, Social core = node size 3, Sex/substances = node size 2, Sex/substances alters only = node size 1, NB: colors used to differentiate ego networks.

majority of alters (98.9%). A total of 63.4% of alters were reported to live in Chicago and provided data sufficient to determine the community area within the city. Alters living in the city were approximately evenly split in the predominant areas with 35% on the north side, 26.2% on the south side and 35.5% on the west side of the city with very few residing in the city center (3.3%).

3.3 Sex and Substance-Using Behavior with Alters

Respondents reporting using substances on average approximately 1-2 times in the last 6 months with 1,914 alters (56.5% of all ties, including sex/drug only and weak ties). The most frequently used substances included alcohol (41.2%), followed by marijuana (22.5%). In terms of sexual behavior, respondents reported having sex with 837 alters (24.7% of all ties, including sex/drug only and weak ties); most frequently male partners (92%). Only 56.0% reported always using a condom during anal sex with male partners in the prior 6 months.

3.4 Reliability of Sociogram vs. Matrix Elicitation of Alters

In our reliability analysis, we found that the coefficient of agreement varied for social network ties in comparison to sex and substance use ties, but all estimates were substantial or nearly perfect according to common criteria (Landis & Koch, 1977), with overlapping confidence intervals for the social and substance use network reliability, but significantly different reliability for the sexual networks. The overall kappa for all ties was 0.74 (95% CI: 0.69, 0.80), whereas for social network ties it was 0.73 (95% CI: 0.66, 0.80), for substance use ties it was 0.68 (95% CI: 0.58, 0.77), and for sex ties it was 0.95 (0.86, 1.0). Reliability was not related to size of any of the networks (i.e., social, substance use or sexual) or overall network size (r=.039, p=.74). Of note, there are two ways our participant-aided sociogram approach could produce results different from the gold standard matrix-based approach: by producing false negatives or false positives. Our method was slightly more likely to produce false negatives than false positives, i.e., among 1,560 observations (i.e., in the sub-sample in which comparison between methods was analyzed); a failure to report a tie (i.e., in comparison to matrix-based elicitation) was observed in 46 cases and in 42 cases our method resulted in a reported link not reported in the matrix approach. Most false positives were reported among social network ties, whereas most false negatives were reported among substance-use ties.

Table 1: Name Generator, Alter Distribution, and Relationship Characteristics Among 175 Ego-Networks of Young Men Who Have Sex with Men

Entire Sample	N	%
Core Network	2579	76.1%
Substance-use or Sex only ties	689	20.3%
Substance-use or Sex only ties to ≥2 alters (not ego)	122	3.6%
Core Network Only	**N**	**%**
Name generator:		
Closest to/talk to/share feelings	1008	39.1%
Time/energy to help	645	25.0%
Lend or give you $25	339	13.1%
Help for gay-related issues	262	10.2%
Spend time with, not as close	325	12.6%
Specificity of Name Information:		
First and last names	2401	93.1%
First name and first initial of last name	22	0.9%
First name only	129	5.0%
Nickname only	27	1.0%
No name-based information	0	0.0%
Type of relationship:		
Immediate family/relative	614	23.8%
Friend	1547	60.0%
School personnel	37	1.4%
Minister/church official	3	0.1%
Community organization staff	18	0.7%
Co-worker	59	2.3%
Acquaintance/associate	121	4.7%
Other	180	7.0%
Strength of relationship (mean)	17	
(1) Very close	1249	48.5%
(2) Somewhat close	967	37.5%
(3) Not at all close	361	14.0%
Frequency of Communication, last 6 months (mean)	3.7	
(0) Not at all	36	1.4%
(1) Once or twice	170	6.6%
(2) Three to six times	214	8.3%
(3) At least a couple of times a month	507	19.7%
(4) Weekly	825	32.0%
(5) Daily	826	32.0%

4. Discussion

In this study, we sought to adapt and test a participant-aided sociogram approach for the study of the social, sexual, and substance use networks of YMSM; to assess the feasibility of data collection using this approach; and to describe personal network characteristics of the target population. In terms of feasibility, we found that potential participants were largely interested in participation (86% agreed to participate and completed interviews) and able to report first and last names of alters (93.1%) as well as sensitive behavior within networks, including substance use and sexual behavior. Only two participants who enrolled in the study later withdrew due to privacy concerns. It should be noted, however, that study participants were already enrolled in the larger parent study and were familiar with the study staff, which likely helped to overcome concerns about providing sensitive information about other individuals in their network.

We found the data collected via the sociogram approach to be of substantial reliability in comparison to a traditional matrix based approach, with an overall coefficient of agreement of .74 for all networks (Landis & Koch, 1977). Reporting of substance use ties in networks was similarly reliable at kappa =.68, while sexual tie reporting was nearly perfect at kappa=.95. The high reliability of sexual ties versus other types of ties may be due to the clear connection reflected in sexual relations versus the more "fuzzy" connection reflected in social ties. The nearly perfect reliability of sexual ties also suggests participants have good memory for reporting sex between alters in their network, and as such they may have a relatively high degree of certainty about these sexual interactions in which they were not directly involved. Similarly to McCarty and colleagues (2007), we found that the freestyle method of eliciting ties in our participant-aided sociogram method resulted in "false negatives," that is, the failure to report a tie in comparison to the matrix-based elicitation; however this was fairly rare (i.e., 46 of 1,560 observations or 3%). We also found that the sociogram approach resulted in "false positives," that is when a tie was reported via the sociogram, but not in the matrix-based approach (i.e., 42 of 1,560 observations, or 2.7%). The false positives were more common for social ties whereas the false negatives were more common for substance use ties. In this case, the pattern may be less the result of social structural patterns (i.e., as McCarty and colleagues concluded with regard to their sample) and more the result of the relative stigma associated with substance use versus social support, resulting of under-reporting of substance use ties and over-reporting of social ties in a freestyle approach. The overall excellent reliability across types of ties suggests that the more efficient and less burdensome sociogram approach for measuring network structure is a viable alternative to asking participants to report on all ties between alters, especially in light of our recent advances in automating the data conversions to visualization software.

In terms of network size and distribution of alters, social networks were comprised of approximately 15 alters on average, although up to 40 could be listed (only

Table 2. Core Alter Characteristics Among 175 Ego-networks of Young Men Who Have Sex with Men

	N	%
Age (Mean; Range)	25.64	4 - 84
≤15	28	1.1%
16-18	344	13.3%
19-22	1194	46.3%
23-29	462	17.9%
30-39	216	8.4%
≥40	310	12.0%
Not reported or Unknown	25	1.0%
Race		
African American	1161	45.0%
Latino	569	22.1%
White	614	23.8%
Other	229	8.9%
Not reported or Unknown	6	0.2%
Sex		
Male	1307	50.7%
Female	1222	47.4%
Transgender	48	1.9%
Not reported or Unknown	2	0.1%
Sexual Orientation		
Bisexual	271	10.5%
Gay	950	36.8%
Queer	35	1.4%
Heterosexual	1283	49.7%
Not reported or Unknown	40	1.6%
Residential Location		
Chicago	1634	63.4%
North Side	572	35.0%
Center	54	3.3%
South Side	428	26.2%
West Side	580	35.5%
Illinois (including Chicago)	2306	89.4%
Out-of-State	245	9.5%
Not Reported or Unknown	28	1.1%

one participant listed all 40), suggesting that limiting social network member names generated to 20, as was done in prior network studies of homeless youth (Kennedy et al., 2012; Tucker et al., 2012) may be a viable option and would further economize the interview process. Approximately 10% of alter names were generated by using a gay-specific name generator, which suggests that a tailored approach to name generators may be important for this population, in particular. Social networks included primarily friends and to a lesser extent family members to whom respondents reported to be quite close emotionally (i.e., very close to somewhat close on average), but with whom they communicated with moderate frequency (i.e., weekly to a couple of times per month on average).

In terms of age and race/ethnicity, alters mirrored egos to a large extent. The majority of social network alters were ages 19-29, with very few alters ages 18 and under. This is an important finding and suggests that many school-age YMSM, i.e., those under age 19 in particular, may be isolated from networks of young adult MSM. This also suggests that studies seeking to recruit adolescent MSM should not rely on young adults for network-based recruitment as there appears to be little bridging across these two developmental groups. Racial homophily was high among all youth with the greatest degree of homophily among African American youth. In terms of gender, approximately half of alters were male and just under half were heterosexual, although a large percentage were reported to be gay (36.8%), as might be expected in a study of YMSM. Given the location of gay ghettos and gay-friendly neighborhoods on Chicago's north side, the fact that alters came from all areas of the city suggests diverse geographic distribution of networks, beyond the most obvious neighborhoods, despite the high level of racial homophily among African America youth in particular.

Substance use between ego and alters was relatively common with egos reporting use of substances with over half of alters, although those substances were largely restricted to alcohol and marijuana and frequency of use was relatively low (i.e., 1-2 times per month on average). Sexual risk behavior between egos and alters was also common, with over half of respondents indicating that they did not use condoms with anal sex partners in the prior 6 months (see Birkett et al., in press and Mustanski et al., 2014 for further description of network factors related to sexual risk in the sample). It is important to note that close to an additional 4 substance-using and sexual ties were identified outside of the social network, highlighting the lack of complete overlap in these sets of ties.

While relative size of personal networks and their composition was not particularly surprising, we are not aware of similar studies that have attempted to estimate the size and characteristics of these networks among YMSM. An additional strength of this study and an extension of prior work is the efficient collection of information on the characteristics and structure of multiple networks (social, substance use and sexual) via this participant-aided visualization approach. While such analyses are beyond the scope of this paper, these data will allow for future analyses of the role of overlapping components and ties on health outcomes of interest.

It is our hope that this study will serve as a reference and resource upon which to build future personal network studies. Our experience demonstrates not only the feasibility of this approach for YMSM, but also the ability to collect large amounts of network data on a population at high risk of HIV infection. While we did not measure satisfaction with the research process, data collection staff reported spontaneous feedback from participants indicating a high level of interest and engagement in the data collection process, particularly with the personal network elicitation on the whiteboard. In addition, staff found this approach straightforward and easy to implement and enjoyed the dynamic interaction with participants. Our experience with this population is consistent with the developmental literature which suggests that the ability to focus, particularly for long periods of time and the ability for complex thinking is still developing (Hunter & Sparrow, 2012), thus methods to make network data collection efficient, while maintaining validity and reliability, are quite important.

In terms of limitations, all network data was collected in interviewer-administered format, therefore social desirability bias may have been a factor in reporting, particularly reporting of sensitive behaviors. In addition, information on network alters was collected via report by the ego and thus is subject to potential errors associated with proxy reporting. While we sought to maintain the low technology approach of Hogan and colleagues (2007) to increase the generalizability of this method to low resource environments, we recognize that the use of whiteboards may not be easily portable. Therefore, this method, while quite appropriate for remote locations, may not be generalizable when portability is called for in field-based work. Data entry and management was perhaps the most challenging aspect, as this approach relied on paper and whiteboard capture initially, with subsequent computer data entry and verification. In future research on participant-aided visualization approaches, investigators should consider assessing low technology approaches such as those described here in comparison to data capture and visual feedback via electronic

devices and software such as EgoNet (McCarty, 2011) and VennMaker (Kronenwett & Duwaerts, 2014), which automate electronic data collection and visualization, for feasibility and satisfaction as well as reliability and validity, particularly in low resource settings where HIV-related research is often conducted.

References

Anderson, R.M., Gupta, S., & Ng, W. (1990). The significance of sexual partner contact networks for the transmission dynamics of HIV *Journal of Acquired Immunodeficiency Syndrome*, 3(4), 417–429.

Aral, S. (1999). Sexual network patterns as determinants of STD rates: Paradigm shift in the behavioral epidemiology of STDs dade visible. *Sexually Transmitted Diseases*, 26(5), 262-264.

Auerswald, C.L., Muth, S.Q., Brown, B., Padian, N., & Ellen, J.M. (2006). Does partner selection contribute to sex differences in sexually transmitted infection rates among African American adolescents in San Francisco. *Sexually Transmitted Diseases*, 33(8), 480-484.

Berkman, LF, & Glass, T. (2000). Social integration, social networks, social support, and health. In L. Berkman & I. Kawachi (Eds.), *Social Epidemiology*. New York: Oxford University Press.

Birkett, M., Kuhns, L. M., Latkin, C., Muth, S.Q., & Mustanski, B. (in press). The sexual networks of racially diverse young men who have sex with men. *Archives of Sexual Behavior*.

Borgatti, S.P. (2002). NetDraw software for network visualization. Lexington, KY: Analytic Technologies.

Centers for Disease Control and Prevention. (2013a). HIV surveillance report 2011 (Vol. 23). Atlanta, GA: Centers for Disease Control and Prevention.

Centers for Disease Control and Prevention. (2013b). HIV among Black/African American gay, bisexual and other men who have sex with men *Fact Sheet*.

Christley, R.M., Pinchbeck, G.L., Bowers, R.G., Clancy, D., French, N.P., Bennett, R., & Turner, J. (2005). Infection in social networks: Using network analysis to identify high-risk individuals. *American Journal of Epidemiology*, 162(10), 1024–1031.

Clatts, Michael C., Goldsamt, Lloyd, Neaigus, Alan, & Welle, Dorinda L. (2003). The social course of drug injection and sexual activity among YMSM and other high-risk youth: an agenda for future research. *Journal of Urban Health*, 80(4 Suppl 3), iii26-39.

Clerkin, E. M., Newcomb, M. E., & Mustanski, B. (2011). Unpacking the racial disparity in HIV rates: the effect of race on risky sexual behavior among Black young men who have sex with men (YMSM). *Journal of Behavioral Medicine*, 34(4), 237-243. doi: 10.1007/s10865-010-9306-4

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Dennis, A. M., Murillo, W., Hernandez, F. D. M., Guardado, M. E., Nieto, A. I., Lorenzana de Rivera, I., . . . Paz-Bailey, G. (2013). Social network-based recruitment successfully reveals HIV-1 transmission networks among high-risk individuals in El Salvador. *Journal of Acquired Immune Deficiency Syndromes*, 63(1), 135-141.

Freeman, L.C., Romney, A.K., & Freeman, S.C. (1987). Cognitive structure and informant accuracy. *American Anthropologist*, 89(2), 310-325.

Heckathorn, D.D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(2), 174-199.

Heckathorn, D.D. (2002). Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49(1), 11-34.

Hogan, B., Carrasco, J.A., & Wellman, B. (2007). Visualizing Personal Networks: Working with Participant-aided Sociograms. *Field Methods*, 19(2), 116-144.

Hunter, Scott J., & Sparrow, Elizabeth P. (2012). *Executive function and disfunction: Identification, Assessment and Treatment*: Cambridge University Press.

Iguchi, M. Y., Ober, A. J., Berry, S. H., Fain, T., Heckathorn, D. D., Gorbach, P. M., . . . Zule, W. A. (2009). Simultaneous recruitment of drug users and men who have sex with men in the United States and Russia using respondent-driven sampling: Sampling methods and implications. *Journal of Urban Health,* 86, S5-S31. doi: 10.1007/s11524-009-9365-4

Kapadia, F., Siconolfi, D. E., Barton, S., Olivieri, B., Lombardo, L., & Halkitis, P. N. (2013). Social support network characteristics and sexual risk taking among a racially/ethnically diverse sample of young, urban men who have sex with men. *AIDS Behav*, 17(5), 1819-1828. doi: 10.1007/

s10461-013-0468-2

Kennedy, DP, Tucker, JS, Green, HD, Golinelli, D, & Ewing, B. (2012). Unprotected sex of homeless youth: Results form a multilevel analysis of individual, social network, and relationship factors. *AIDS Behav*, 16(7), 2015-2032.

Kronenwett, M., & Duwaerts, T-S. (2014). VennMaker (Version 15) [Visual network mapping software]: Kronenwett & Adolphs - Tools and Services for Social Network Analysis. Retrieved from www.vennmaker.com

Kuhns, L. M., Kwon, S., Ryan, D. T., Garofalo, R., Phillips, G., 2nd, & Mustanski, B. S. (2014). Evaluation of respondent-driven sampling in a study of urban young men who have sex with men. *J Urban Health*, 92(1), 151-167.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

Latkin, C.A., & Knowlton, A. R. (2005). Micro-social structural approaches to HIV prevention: a social ecological perspective. *AIDS Care*, 17(Supplement 1), S102-S113.

Latkin, CA, Forman, V, Knowlton, A, & Sherman, S. (2003). Norms, social networks, and HIV-related risk behaviors among urban disadvantaged drug users. *Social Science & Medicine*, 56(3), 456-476.

Laumann, E.O., Galaskiewicz, J., & Marsden, P.V. (1978). Community structure as Interorganizational linkages. *Annual Review of Sociology*, 4, 455-484.

Marsden, PV. (1990). Network data and measurement. *Annual Review of Sociology*, 16, 435-463.

McCarty, C. (2011). EgoNet (Version GPLv2) [Visual network mapping software]: University of Florida. Software. Retrieved from www.sourceforce.net

McCarty, C., Molina, J.L., Aguilar, C., & Rota, L. (2007). A comparison of social network mapping and personal network visualization. Field Methods, 19(2), 145-162.

Moreno, JL. (1953). *Who shall survive? : Foundations of sociometry, group psychotherapy, and sociodrama* (2nd ed.). Beacon, NY: Beacon House.

Morris, M., & Kretzschmar, M. (1997). Concurrent partnerships and the spread of HIV. *AIDS*, 11(5), 641-648.

Mustanski, B., Birkett, M., Kuhns, L. M., Latkin, C., & Muth, S.Q. (2014). The role of geographic and network factors in racial disparities in HIV among young men who have sex with men: An egocentric network study. *AIDS and Behavior*. Advance online publication.

Mustanski, B., Newcomb, M. E., & Clerkin, E. M. (2011). Relationship characteristics and sexual risk-taking in young men who have sex with men. *Health Psychology*, 30(5), 597-605.

Mustanski, B.S., Newcomb, M. E., Du Bois, S.N., Garcia, S.C., & Grov, C. (2011). HIV in young men who have sex with men: A review of epidemiology, risk and protective factors, and interventions. *Journal of Sex Research*, 48(2), 218-253.

Newcomb, M. E., & Mustanski, B. (2013). Racial differences in same-race partnering and the effects of sexual partnership characteristics on HIV risk in MSM: A prospective sexual diary study. *Journal of Acquired Immune Deficiency Syndromes*, 62(3), 329-333.

Nunnally, JC, & Bernstein, IH. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill, Inc.

Potterat, JJ, Rothenberg, RB, & Muth, SQ. (1999). Network structural dynamics and infectious disease propagation. *International Journal of Std & Aids*, 10, 182-185.

Potterat, JJ, Woodhouse, DE, Muth, SQ, & et al. (2004). Network dynamism: History and lessons of the Colorado Springs study. In M. Morris (Ed.), *Network epidemiology: A handbook for survey design and data collection* (pp. 87-114). New York: Oxford University Press Inc.

Ramirez-Valles, J., Heckathorn, D. D., Vazquez, R., Diaz, R. M., & Campbell, R. T. (2005). From networks to populations: The development and application of respondent-driven sampling among IDUs and Latino gay men. *Aids and Behavior*, 9(4), 387-402. doi: 10.1007/s10461-005-9012-3

Reisner, S. L., Mimiaga, M. J., Bland, S., Skeer, M., Cranston, K., Isenberg, D., . . . Mayer, K. H. (2010). Problematic alcohol use and HIV risk among Black men who have sex with men in Massachusetts. *Aids Care*, 22(5), 577-587.

Rhodes, S. D., McCoy, T. P., Hergenrather, K. C., Vissman, A. T., Wolfson, M., Alonzo, J., . . . Eng, E. (2012). Prevalence estimates of health risk behaviors of immigrant Latino men who have sex with men. *Journal of Rural Health*, 28(1), 73-83.

Rice, E., Barman-Achikari, A., Milburn, N. G., & Monro, W. (2012). Position-specific HIV risk in a large network of homeless youths. *American Journal of Public Health*, 102(1), 141-147.

Rothenberg, R. , & Muth, S.Q. . (2007). Large-network

concepts and small-network characteristics: Fixed and variable factors. *Sexually Transmitted Diseases*, 34(8), 604-612.

Rothenberg, R., Baldwin, B., Trotter, R., & Muth, S.Q. (2001). The risk environment for HIV transmission:results from the Atlanta and Flagstaff network studies. *Journal of Urban Health*, 78(3), 419-432.

Schneider, J., Michaels, S., & Bouris, A. (2012). Family network proportion and HIV risk among Black men who have sex with men. *Journal of Acquired Immune Deficiency Syndromes*, 61(5), 627-635.

Smith, Kirsten P., & Christakis, Nicholas A. (2008). Social networks and health. *Annual Review of Sociology*, 34, 405-429.

Staras, SA, Cook, RL, & Clark, DB. (2009). Sexual partner characteristics and sexually transmitted diseases among adolescents and young adults. *Sexually Transmitted Diseases*, 36(4), 232-237.

Tucker, JS, Hu, Jianhui, Golinelli, D, Kennedy, DP, Green, HD, & Wenzel, SL. (2012). Social network and individual correlates of sexual risk behavior among homeless MSM youth. *J Adolesc Health*, 51(4), 386-392.

Wasserman, S, & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.

Wohlfeiler, D., & Potterat, JJ. (2005). Using gay men's sexual networks to reduce sexually transmitted disease (STD)/human immunodeficiency virus (HIV) transmission. *Sexually Transmitted Diseases*, suppl 32(10), s48-s52.

The Content Structure of Intelligence Reports

Kimmo Elo
University of Turku
Turku, Finland

Abstract

Despite its close connection to many of the methodological questions and problems related to uncloaking hidden structures or to analyzing network dynamics tackled by network analysts in tterrorism or crime studies, researchers on Cold War intelligence have shown limited interest in network analysis. Although there might be material-related reasons for that, the resistance is to a great extent caused by the unfamiliarity with the method itself. Following the idea that how one looks at the material determines what they see, this article will evidence how social network analysis could be applied to historical sources in order to extract, analyze and visualize new knowledge. The analysis will illustrate the capability of SNA-based keyword co-occurrence network analysis and visualizations for uncovering and identifying the corpus' structural properties, detecting the thematic backbone of the report corpus, and for analyzing network dynamics. The material used in the analysis consists of reports on Nordic affairs produced in 1975-1989 by the East German foreign intelligence service. Besides keyword co-occurrence network analysis and visualizations, the article shows how community detection techniques can be used to extract thematic backbones in a report corpus. A critical assessment of the results obtained by SNA techniques in their historical context confirms their validity and reliability. More generally, the results showed that combining SNA with methods of content analysis offers a promising perspective for developing new research methods for the analysis of the social network of language capable of tackling and extracting contextual and semantic relationships in networks based on textual data. Furthermore, the SNA-based approach to textual networks could open up new perspectives for exploratory historical research with the view to finding new foci for research and comprehending new research hypothesis and questions.

Keywords: SNA, Co-occurrence networks, Content analysis, Network communities, Visualizations, East German foreign intelligence, Intelligence reports, Cold War

Authors
***Kimmo Elo**,* is an adjunct professor and senior lecturer in the Institute for Political Science and Contemporary History at the University of Turku in Turku, Finland and The Institute for German Studies at the Åbo Akademi University in Turku, Finland. His research interest include Historical network research, German politics and history, Intelligence studies, Cold war history, European integration and Digital Humanities.

Correspondence concerning this work should be addressed to Kimmo Elo, Department of Political Science and Contemporary History, FIN-20014 University of Turku, Turku, Finland. Email: kimmo.elo@utu.fi.

DOI: http://dx.doi.org/10.17266/35.1.2

1. Introduction

The very essence of all intelligence services is related to their networking capacity. A closer look at the so-called intelligence cycle, the cyclical process consisting of information-gathering (single-source collection), exploitation, (all-source) analysis and dissemination[2], quickly reveals how fundamentally important networking capabilities and networks are for all intelligence services. First, each intelligence service attempts to build a functioning, dense network of human (HUMINT[3]) or non-human (SIGINT[4]) sources providing the service with raw data. In intelligence analysis, analysts combine information from the collected raw data with previous knowledge and information from other sources in thematic reports. From this perspective an intelligence report consists of logically connected pieces of information, i.e. an information network. Finally, these reports are disseminated to a small or large network of recipients at various levels of governmental administration.

Taking into account the importance of intelligence during the Cold War and the centrality of networks and networking (both human and non-human) for intelligence services, the limited interest in network analysis among Cold War studies is somewhat surprising. One reason might be that Cold War studies are still dominated by historians who are neither primarily interested in intelligence studies[5] nor familiar with network analysis as a research method[6]. Conversely, recent research in terrorism or crime studies has discussed empirical, methodological and theoretical questions and problems in relation to uncovering hidden structures or analyzing network dynamics[7], which are also relevant for intelligence studies. But, outside the humanities, the opposite holds true as network studies have shown only modest interest in historical sources. There may be a rational explanation for this lack of interest. If empirical data is used just for algorithm testing, the use of materials requiring in-depth knowledge of source criticism and time-consuming preparation simply makes no sense.

Despite its close empirical connection with the Cold War period, this article intends to contribute to methodological and theoretical discussions in intelligence studies. Following the idea that research methods have a strong impact on what can be achieved, this article will show how network analysis could be applied to historical sources in order to extract and visualize new knowledge. The article focuses on two questions. First, how can social network analysis (SNA) be applied in order to study the characteristics of co-occurrence networks of keywords summarizing the content of intelligence reports? Second, how can the meta-data of intelligence reports be used for tackling characteristics and dynamics of the report corpus? The analysis will also illustrate the capability of clustering techniques for uncloaking and identifying the corpus' structural properties, detecting the thematic backbone of the report corpus, and for analyzing network dynamics.

The structure of this article is as follows. Section two discusses the research method and material used. On this basis, the article will introduce a technique combining co-occurrence analysis and a method for identifying the key concepts in a document collection functioning as junctions of different documents. In regard to material, the posthumous data from the HV A archive available for research will be introduced. Section three consists of analysis and visualizations focusing on content structures, dynamics and thematic clusters in the keyword co-occurrence networks. The article will conclude with a critical assessment of the most important findings in the broader context of Digital Humanities.

2. Method and Material

2.1 Data Analysis and Visualization Methodology

The idea of uncloaking hidden structures from texts with the help of graphical tools and representation is not new. Since the late 1990s, humanists and social scientists have shown an increasing amount of interest on network analysis, a development, which has resulted in the emergence of new concepts and methods[8], including text mining, semantic network analysis and content analysis, which are currently widely used by humanists focusing on structure-oriented analysis of large text materials.

The concept of co-occurrence networks is a

[2] Herman, 2001, 79; Bruce & George, 2008, 2; Walsh, 2011
[3] HUMINT stands for "human intelligence", i.e. for intelligence-gathering by interpersonal contacts.
[4] SIGINT stands for "signal intelligence", i.e. intelligence-gathering by intercepting all kind of communication or electronic signals.
[5] Garthoff, 2004, 21
[6] Although network analysis has been used in a wide range of historical case studies (an up-to-date bibliography of historical network research can be found at: https://www.zotero.org/groups/historical_network_research_bibliography/items [on-line: visited on August 10, 2014]), to our knowledge it has not been applied to historical intelligence studies.
[7] E.g. Krebs, 2002; Raab & Milward, 2003; Xu & Chen, 2005; Enders & Su, 2007; Schwartz & (D.A.) Rouselle, 2009; Hutchins & Benham-Hutchins, 2010; Malm & Bichler, 2011; Morselli, 2010
[8] For a recent review, see Schultz-Jones, 2009

powerful method for the graphical representation of potential and existing relations between different concepts, terms or other entries in textual materials. Co-occurrence networks are widely used among scholars interested in content analysis or text mining through which an understanding of the thematic structure of a text corpus is sought.[9] In recent studies, co-occurrence network based approaches have been applied in constructing semantic networks of scientific journals[10], extracting social, issue or concept networks from large datasets[11], analyzing customer feedback or assessing free text answers[12], and in creating semantic summaries of documents.[13]

Notwithstanding its growing popularity among scholars, co-occurrence network analysis has mainly been used to examine and visualize relations between concepts in order to find out the most or the least connected concepts, semantic clusters within a text corpus or changes over time in co-occurrences. Scholars have shown less interest in the network structure itself, its characteristics and network metrics. Although most studies on co-occurrence networks have several graphical representations, the visualizations are rarely discussed, let alone analyzed. In most papers visualization settings (layout used for visualization, layout settings etc.) and their impact on the graphical representations are simply left unexplained. One central reason for this lack of interest in network structure and characteristics is the fact that most co-occurrence analyses have focused on the interconnections of concepts, not their interaction or influence on the network structure as a whole. In other words, in most co-occurrence analyses visualizations serve as graphical representation of the content in the form of paired concepts.

Mastering the shift from the analysis of simple concept interconnections to concept co-occurrence networks requires a methodological bridge-building from co-occurrence analysis to SNA. The analysis and visualizations conducted in this article are based on Paranyushkin's (2011) recent work on using SNA for identifying meaning circulation in text documents as well as on Feicheng & Yating's (2014) study of the use of SNA for co-occurrence network analysis of on-line tags. Both papers share the relatively simple, yet powerful understanding that SNA - originally designed for analysis of social or human relationships - could be applied to co-occurrence networks of textual data as well. Although this slightly changes the terminology "actors" are replaced by "concepts" or "keywords" - it does not affect the focus; the emphasis remains on exploring and understanding the structures and internal relationships among concepts / keywords, not just on presenting the connections between them. Both papers also powerfully exemplify how SNA methods can be used to extract and visualize the basic characteristics and the structure of the co-occurrence network. This article seeks to illustrate how network metrics can be used for comparisons between different time periods in order to uncover network dynamics.[14]

The method for generating co-occurrence networks of report keywords is based on the understanding that each report forms a particular thematic space which is summarized in keywords. The underlying assumption is that keywords are used to describe content in a similar way than tags are used for tagging on-line documents. Consequently, if one creates keyword co-occurrence matrices for each report in the report corpus, it can be assumed that thematically comparable reports will have an analogous keyword co-occurrence structure. Moreover, because the East German foreign intelligence service used a standardized set of keywords for report tagging and the keywords were added according to the principle of relevance, we are expecting to identify core keywords constructing the core knowledge of the report corpus, i.e. keyword co-occurrences linking different reports to each other. To keep the order of relevance, the co-occurrence matrix is constructed by pairing the first keyword (which is considered as the most relevant) with all subsequent keywords. Based on this method, a network of keywords and their co-occurrences - G(V, E) - is built, where V is the set of keywords and E is the set of edges (co-occurrences).[15]

2.2 Intelligence Report Data of East German Foreign Intelligence

The East German foreign intelligence service's main department A (*Hauptverwaltung A*, abbr. HV A) was formally a part of the East German State Security Service (*Ministerium für Staatssicherheit*, abbr. MfS, Stasi). Like all intelligence services, the HV A was responsible for running the complete intelligence cycle from information

9 E.g. Stuart & Botella, 2009; Lee et al., 2010; Brier & Hopp, 2011
10 E.g.Stuart & Botella, 2009; Hsu & Kao, 2013
11 E.g. Diesner et al., 2012; Holt et al., 2012; Novotny & Cheshire, 2012; Feicheng & Yating, 2014
12 Eklund et al., 2011; Noorbehbahani & Kardan, 2011
13 Özgür et al., 2008; Oesper et al., 2011; Paranyushkin, 2011; Yang et al., 2014
14 See also Paranyushkin, 2011, 20-21; Marres, 2012, 158
15 See also Feicheng & Yating, 2014, 234-235

gathering to report dissemination according to objectives and guidelines decided by the party and state leadership[16], i.e. ”torn between its twin skills of collecting information and evaluating it”.[17] The main task of the HV A was to gather, evaluate and process technical, scientific, military and political information. It was also responsible for compiling reports on political, economic and strategic issues, disseminated not only to responsible state and party organizations inside the GDR, but also to allied services inside the Soviet Empire, most importantly to the Soviet KGB. Put in theoretical terms, the HV A was a huge information system responsible for collecting, storing, analyzing and disseminating information. Its systemic structure was built upon a center-periphery schema: the periphery was responsible for gathering intelligence while the role of the center was to control, direct and supervise activities in the periphery as well as to process the collected information to reports for dissemination.[18] The backbone of this information system formed an extensive network of ”unofficial collaborators”, the IMs (*Inoffizielle Mitarbeiter*), organized in HUMINT networks operating in all Western European countries. In 1989, approximately 189.000 IMs operated for the Stasi and about 15.000 of them for the HV A.[19]

In 1990, the archive of the HV A was almost completely destroyed. The remaining materials are maintained by the Agency of the Federal Commissioner for the Stasi records (*Der Bundesbeauftragte für die Unterlagen des Staatssicherheitsdienstes der ehemaligen Deutschen Demokratischen Republik*, abbr. BStU). In total, the Agency maintains approximately 111 kilometers of archived files of which the share of the HV A is just 47 meters.[20] However, in 1998 experts of the BStU succeeded in decrypting an operational database system of the HV A called SIRA (System der Informationsrecherche der HVA). This system was instigated in the mid-1970s and used to maintain the intelligence cycle, administer undercover operations, and most importantly, to store meta-data of information and reports.[21] Because SIRA was developed for the maintenance of daily information flows to and from the HV A, it opens a window into the HV A’s daily operational work and provides scholars with operational data related to the intelligence cycle from the perspective of the HV A. The SIRA entries cover the years from 1969 to 1989, but the completeness of the stored information varies a great deal. In addition, SIRA records do not contain original documents, and thus, do not substitute the destroyed archival files. Rather, the SIRA records are comparable to bibliographical entries in a library catalog giving the user information

Figure 1: HV A reports on Finnish affairs 1975-1989. (Source: Author’s calculations based on material from BStU.)

16 See Müller-Enbergs, 1998, 40-41; Müller-Enbergs, 2008, 5
17 Herman, 2001, 3-4
18 On information systems, see e.g. Avison & Myers, 1995; Alter, 2008; Hyvönen et al., 2008
19 Müller-Enbergs, 2011, 21
20 Müller-Enbergs, 2011, 11-12
21 SIRA was built as a relational database and consists of four main tables (sub-databases) numbered from 11 to 14. Each sub-database is dedicated to a specific domain of intelligence: ”Scientific and technical espionage” (sub-database #11), ”Problems and operations related to domestic and foreign policies, economy and military politics outside the GD” (sub-database #12), ”Political relations in the operation area” (sub-database #13) and ”Counter-intelligence” (sub-database #14). Additionally, administrative information of operations - e.g. supervisor changes, opening of new files - are stored in the sub-database #21. (Konopatzky, 2003; Müller-Enbergs, 2007, 13ff.)

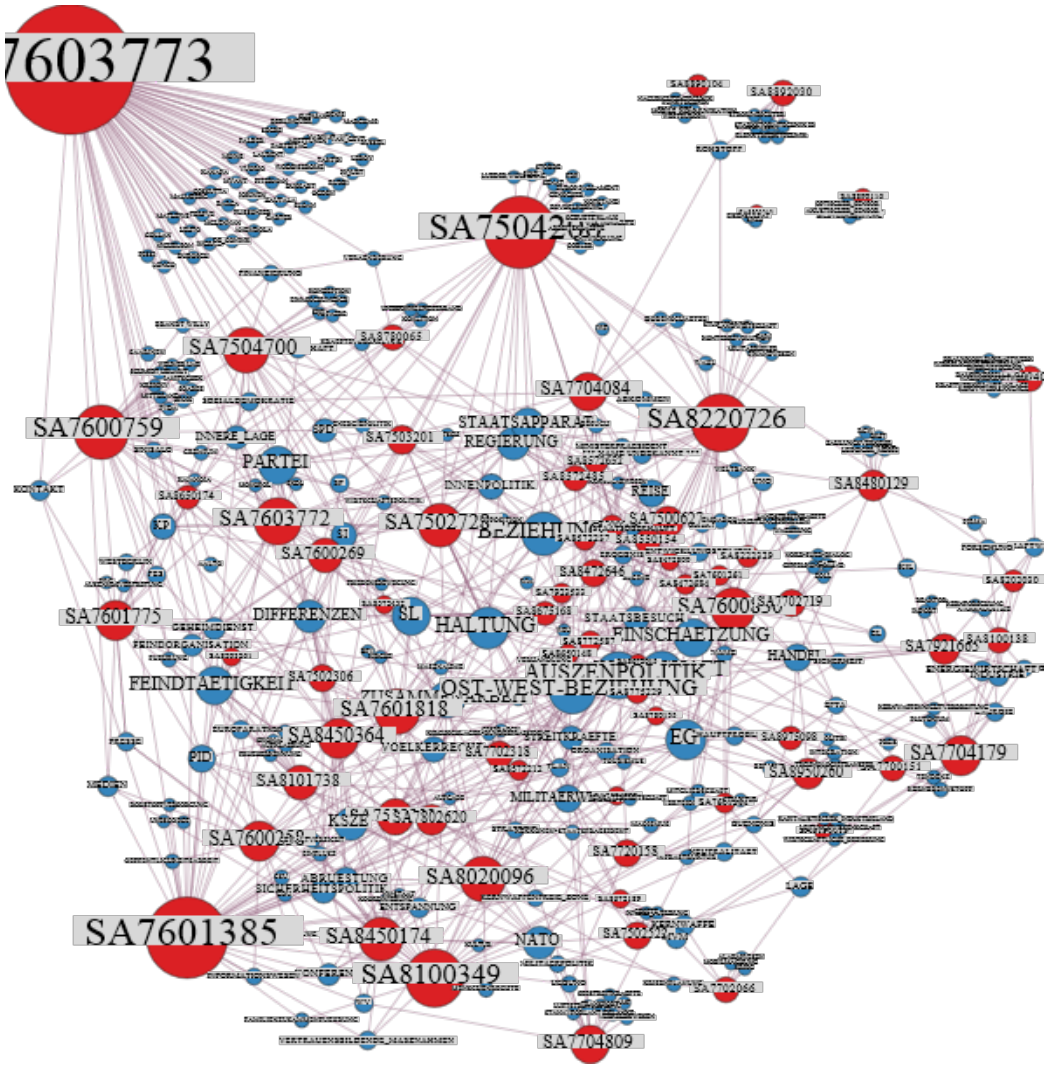


Figure 2: Keywords per report network.

of the collections. In the case of the HV A, however, the collections have been destroyed and only the catalog exists.[22]

The material used in this article consists of a selected corpus of dissemination (type "SA"[23]) records on Finnish and Nordic affairs from 1975 to 1989. During the Cold War, the four Nordic countries - Finland, Sweden, Norway and Denmark - formed an interesting geopolitical area in the European north, characterized by both differences and similarities.[24] As recent studies have shown, the HV A conducted intelligence operations also in the European North. The Nordic countries were an important, but not the central operational area for the HV A and undoubtedly these countries enjoyed a special status for the East German foreign intelligence.[25] The year 1975 is widely regarded as a turning point in the GDR's history, but also in European politics. The main reason for this assessment is the Conference on Security and Cooperation in Europe (CSCE) held in Helsinki in August 1975. During the second half of the 1970s, the political consequences of the CSCE resulted in growing tensions within the Soviet empire, including the GDR. The party leadership in the GDR was increasingly concerned about the destabilizing impact of the CSCE

22 Researchers cannot directly access the database, but SIRA queries are carried out by BStU according to search criteria defined by the researcher. Since the database is administered in a SQL-based system, complex multi-criteria queries are possible. The results are available in printed form only and BStU charges a small per-page fee (currently 10 cents/page).

23 Each SIRA record has an unique ID starting a two-character string ("SE", "SA" or "SB") describing the information type, followed with two digits identifying the recording year and five digits from the database counter. Records marked with "SE" (SIRA Eingang) are input records, i.e. meta-data of intelligence gathered by the HV A. Records marked with "SA" (SIRA Ausgang) contain the meta-data of disseminated material (reports, evaluations etc.) the HV A has disseminated to external partners. Finally, records marked with "SB" (SIRA Bestellung) are records storing meta-data for intelligence requests from outside. As an example, a SIRA record with the ID "SA7503201" is a dissemination record (type: SA) stored in 1975 (SA**75**03201).

Table 1: Force Atlas layout parameters used for visualizations.

Parameter	Value
Repulsion strength	200.0
Attraction strength	10.0
Gravity	30.0
Attraction distribution[a)]	True

[a)] Distributes the attractive force along outbound links, thus pushing hubs at the periphery and putting authorities more central.

on its power and consequently instructed the HV A to conduct continuous evaluations of the situation in Europe.[26] These developments were boosted by Mikhail Gorbachev's rise to power in the Soviet Union in 1985. Against this background it is worth analyzing whether the HV A reports on Nordic affairs also reflect these tectonic changes in Europe. More generally, the analysis seeks to signal two aspects. First, keywords are a valuable source of knowledge capable of providing valuable information about content structures and dynamics. Second, SNA offers well-suited methods for network analysis revealing valuable information about the relationships within the network. A keyword or tag co-occurrence network should be considered as the embodiment of relevant information about the underlying structure and relationships. Here the devil is the detail: two co-occurrence networks may look very similar, but, at the same time, show significant deviation in network metrics. SNA offers solid methods for capturing the concealed, but important differences.

The report corpus used in the analysis consists of 69 reports. Almost a one third of them, 28 reports, were produced in 1975-1977. Another peak in reporting, a total of 18 reports, occurred in 1984-1986 (see Figure 1). The data preparation was carried out in three steps. First, all paper documents were scanned for optical recognition. Second, each document was processed by Tesseract[28] software for optical character recognition (OCR) and stored in text format. Third, a small Python program for processing the text files in order to recognize different fields, extract field contents, and store the extracted data in a database was developed. Although the results from the OCR processing were of relatively good quality, our program was equipped with some learning and correction functionality. The following information was extracted from the meta-data and stored in the database: 1) original date of the report 2) content keywords 3) country references 4) references to objects (institutions, parties, universities, etc.) and persons. Finally, the keyword co-occurrence matrix was created according the methodology described above. Since the original keywords cross-reference to objects and persons, the keyword co-occurrence matrices were constructed in the following manner: if the special German keyword OBJEKT (object) or NAME (name) appeared in the keyword list, it was removed and the subsequent "real" keywords were upgraded. Object and person references were added to the end of the keyword vector. This practice should ensure the priority of thematic keywords in the pairing process. The content-related keywords were considered as more important structural properties because their purpose is to summarize the content of the report.

3. Using SNA to Explore Intelligence Reports

The readability and interpretation of all graphical representations of network data depends on the quality of the visualization layout. Although network analysis and visualization software have taken quantum leaps, different programs continue to produce somewhat different results. The following analysis and visualizations are prepared with Gephi.[29] The layout used in visualizations is based on Force Atlas algorithm, which seeks to push the most connected nodes - the so-called *hubs* - away from each other (in order to visually highlight the different communities in the graph) and to align the nodes connected to a hub in clusters around them. Moreover, the algorithm seeks to position the authorities into the center ground.[30] Thus in the following graphical representations the most important concepts in the report corpus are located in the middle of the graph. The most important layout parameters used for all visualization are listed in Table 1.

The keywords-per-report network visualizes the distribution of different keywords (n=293) along the report corpus (Figure 2). In the graph, each report is presented as a red circle and each keyword as a blue circle. The dense sub-network in the middle indicates that several reports share the same keywords, and thus, it is not possible to identify clear overall allocation of keywords. However, we can isolate some reports, such as SA7603773, SA7504200, SA8890104 and SA8892030

24 Musiał, 2009, 287. See also Steinbock, 2008, 199ff.
25 For a good summary, see Friis et al., 2010
26 E.g. Schroeder, 1998, 233ff.; Gieseke, 2008
27 Brown, 1996; Gaddis, 2005, 229ff.
28 https://code.google.com/p/tesseract-ocr/
29 https://gephi.org/. See also Bastian et al., 2009
30 http://www.slideshare.net/gephi/gephi-tutorial-layouts (slides 11-12) [on-line. Last visited 9th October 2014]

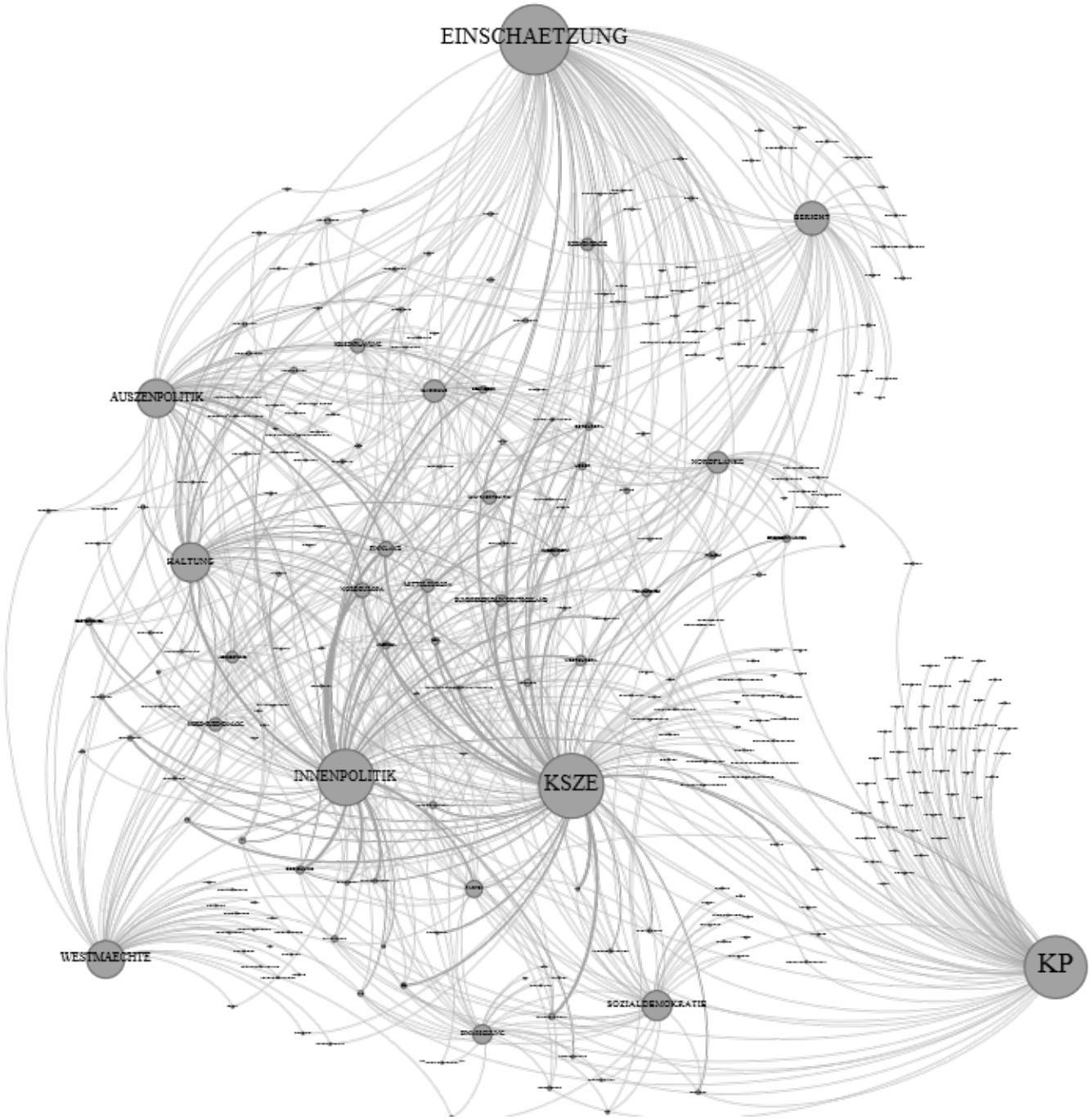


Figure 3: Keyword co-occurrence network 1975-1989.

that share only some keywords with the rest of the corpus. This visual observation is confirmed by network statistics: the network density (the actual proportion of ties in a network) is very low (d=0.012) and the average degree (the number of links in the network compared to number of nodes) is also low (only 2.222). Regarding individual keywords, the top-5 connected keywords (keywords with the highest degree) are: *foreign policy* (deg=20), *east-west relationship* (20), *foreign relations* (18), *economy* (17) and the *European Community* (EC, 16).

Although the keywords-per-document network is useful for identifying core keywords in the corpus, the relationships among the keywords are more interesting in regard to the content structure. Similar to social networks, keyword co-occurrences construct a ”social network of language in which the individuals or actors are not the members of a group, but terms [keywords], and the links are the relationships among them”[31]. Consequently, the significance of the type of interactions that occur between keywords increases in importance. Both the density of co-occurrences (how often two keywords co-occur) and

the number of keywords that relate to the same keyword are significant when assessing the complexity in the network.[32]

The next graph visualizes the keyword co-occurrence network (see Figure 3). In order to visually highlight some characteristics of the network structure, visual effects are added to the graph. First, the size of a node is proportional to its degree, i.e. the more connections the node has the larger it appears in the graph. Consequently, keywords co-occurring with many other keywords can easily be identified in the graph. Second, the thickness of an edge is proportional to its weight, i.e. how many times the co-occurrence occurs in the data. Since the pairing method produces singular co-occurrences at report level, a co-occurrence's weight equals to the number of reports sharing that keyword co-occurrence. Thus, thick edges indicate core keyword co-occurrences in the report corpus. Third, the font size of node labels is proportional to the node's betweenness centrality. In network theory, centrality indicates a node's position in the network and can be calculated either relative to a node's direct neighbors or the whole network. Betweenness, as the term itself indicates, defines centrality by analyzing where a node is placed within the network. Consequently, a node's betweenness centrality score is computed by taking into consideration the rest of the network and by examining how many times a node sits on the shortest path linking two other nodes to each other. Thus, using betweenness centrality as an attribute for the graph helps us to identify nodes that have a "high probability of occurring on a randomly chosen shortest path between two randomly chosen vertices".[33]

In this article, the concept of betweenness centrality is preferred to eigenvector centrality because an attempt is made to uncover thematic pathways in the report corpus. In this respect, keywords functioning as mediators between different clusters / contexts in the report corpus are considered as more significant. A node's eigenvector betweenness is rather reliant on the ties of the node's connections, whereas betweenness centrality depends on the node's capability to act as a connection between two or more nodes that would otherwise remain disconnected. Considering thematic pathways, the latter capability is assumed to be more relevant and therefore betweenness centrality is used to measure and visualize a keyword's status in the keyword co-occurrence network.

In this article, betweenness centrality is interpreted as an indicator for a concept's role in the report corpus in a straightforward manner: nodes with high betweenness centrality metric are considered as mediators between different clusters / contexts found in the whole report corpus, thus revealing the variety of contexts the concepts appear in while concepts with a low(er) betweenness centrality metrics are central in a sub-corpus only. In other words, a node with a high degree but a lower betweenness centrality might be an important local hub in one cluster, but less influential within the whole network because it has only few connections with other clusters. In turn, a node with fewer connections (lower degree), but high betweenness centrality metrics is considered to be adjacent to the most reports in the network.[34]

The overall density of the keyword co-occurrence network is 0.013, which shows only weak connections. The average path length is 3.245, which is relatively short and indicates that each keyword can easily reach another keyword. The clustering coefficient measuring the number of node triangles in a graph is quite low, 0.083. This is due to the size of the network (291 nodes and 582 edges) and the pairing method: suppose one report with keywords A, B and C, resulting in keyword co-occurrences A-B and A-C. Now, a triangle would require another report with the keywords B and C so that either B or C is the first keyword. Actually, the relatively low clustering coefficient metric imply the existence of a set of core keywords forming the thematic backbone of the report corpus.

If this assumption holds true, these keywords should have the highest betweenness centrality values. According to the network data, the three most important keywords include *communist party* (labelled KP in the graph with the betweenness centrality of 13244.091), *CSCE* (KSZE, 10264.730) and *foreign policy* (AUSZENPOLITIK, 9404.719). However, the network graph reveals another story (cf. Figure 3), visualizing foreign policy as the most central concept in the report network. Since the graph is produced with parameters pushing hubs at the periphery and putting authorities more central (see also Table 1), *communist party* seems to be a hub instead. A closer analysis of the report data supports this interpretation: the keyword *communist party* co-occurs 77 times in 8 reports, whereas *CSCE* co-occurs 138 times in 12 reports and is thus a more central keyword than *communist party*. The keyword *foreign policy* has

31 Stuart & Botella, 2009, 15
32 Stuart & Botella, 2009, 15; Feicheng & Yating, 2014, 234
33 Hsu & Kao, 2013. See also Prell, 2012, 103-104
34 See also Paranyushkin, 2011, 13-14; Feicheng & Yating, 2014, 235

Keyword	BC	Deg	Keyword	BC	Deg	Keyword	BC	Deg	Keyword	BC	Deg
Overall (1975-1989)			*1975-1979 (n_r=31)*			*1980-1984 (n_r=15)*			*1985-1989 (n_r=23)*		
Communist party	17844.413	81	Communist party	13581.169	81	CSCE	3837.702	65	Finland	2315.959	16
CSCE	13783.129	95	Evaluation	10601.759	91	Domestic policy	3194.804	43	Foreign policy	1795.726	36
Evaluation	11690.059	92	CSCE	10082.675	84	Foreign policy	2978.139	52	Raw materials	1644.177	28
Foreign policy	11475.587	50	Domestic policy	6262.003	73	Trade	1924.167	33	Energy technology	763.831	10
Domestic policy	10499.800	73	Western powers	4791.409	48	Trust building actions	1548.885	43	Integration	759.899	17
Keyword co-occurrence		N	Keyword co-occurrence		N	Keyword co-occurrence		N	Keyword co-occurrence		N
Overall (1975-1989)			*1975-1979*			*1980-1984*			*1985-1989*		
CSCE↔East-West relations		7	Domestic policy↔Northern Europe		5	CSCE↔EC		2	East-West relations↔Finland		3
CSCE↔Political-ideological diversion		6	Domestic policy↔Finland		5	CSCE↔Finland		2	East-West relations↔GDR		3
CSCE↔Rival actions		6	CSCE↔FRG		4	CSCE↔Northern Europe		2	Raw materials↔FRG		2
CSCE↔EC		5	CSCE↔Northern Europe		4	CSCE↔FRG		2	Raw materials↔Finland		2
CSCE↔Co-operation		5	CSCE↔Central Europe		4	CSCE↔East-West relations		2	Raw materials↔Switzerland		2

Table 2: Dynamics of the top-5 keywords and keyword co-occurances (1975-1989).

the least co-occurrences (74), but appears in 20 reports. In other words, the keyword *foreign policy* has fewer co-occurrences than the other central concepts, but it links more reports to each other. Further, the notion made by Feicheng & Yating (2014, 235) holds true also in our analysis: the keywords having the highest betweenness centrality have low closeness centrality, which measures a node's independence in form of its closeness to other actors and *vice versa*. For example, the keyword *foreign policy* has the 8th *lowest* closeness centrality.

The previous analysis has focused on the overall network topology and represented the report corpus as a static network. However, historical research is mainly interested in changes in time, developments in a certain period of time and the dynamics behind or resulting in or from historical processes. In the recent years, network dynamics has gained in importance and new approaches and methods either to model change in network structures or test explanations for observed change have been constructed.[35] The underlying assumption is that we can capture the dynamics of the report network by slicing the material in snapshots consisting of reports from five-year periods. The network data used in this analysis consists of both the keywords and country references of each report, and thus, allows the tracking down not only of the thematic dynamics, but also possible changes in geographical focus.

To start with the basic network characteristics, the network becomes slightly denser over time: 0.016 (1975-1979), 0.028 (1980-1984) and 0.031 (1985-1989). At the same time, however, the network size decreases first from 296 nodes and 696 edges (1975-1979) to 154/336 (1980-1984) and finally to 112/192 (1985-1989). At the same time, the average path length (2.954 / 2.911 / 3.161) and clustering coefficient (0.102 / 0.093 / 0.057) remain relatively stable. The last period (1985-1989) is quite interesting, since the number of reports is bigger than in 1980-1984, but at the same time the co-occurrence network is smaller and less dense, implying changes in the set of keywords used for summarizing the report contents. A closer analysis of the set of keywords reveals, that the average number of keywords per document changes from 13 (1975-1979) to 12 (1980-1984) and to 6 (1985-1989). At the same time, the total number of keywords used drops from 217 to 99 and then to 83.

These changes in the set of keywords clearly affect the content structure as well. Considering, first, the thematic pathways build around the core keywords, the most remarkable shift occurs in 1985 (Table 2). Until 1985, keywords related to CSCE dominate the co-occurrence network. This content structure well correlates with the historical fact that from the second half of the 1970s onwards the CSCE's political consequences for the Communist camp began to dominate the GDR's

35 For a good summary, see Scott, 2013, 139-145

political agenda. The party leadership in the GDR was increasingly worried about the destabilizing impact of the CSCE on its power and, consequently, instructed the HV A to continuous evaluation of the situation in Europe. Also in the early 1980s, CSCE remained central in the report corpus. However, during this period the CSCE was more and more embedded in the wider context of Western European integration, East-West divide and emerging contacts between socialist and capitalist countries.[36] In 1984, the official visit of the Secretary General Erich Honecker in Finland dominated the reporting of the HV A and most of the reports evaluated both Finnish and West-German reactions before and after Honecker's visit. The most significant change in the content structure occurs from 1985 onwards. First of all, the reports became more focused on Finland and Finland's western relations. But the changes also imply that economic issues and questions related to energy production gained in importance. These changes also correlate well with historical developments: in the late 1980s bottlenecks in the GDR's energy sector worsened dramatically as the Soviet Union cut its oil exports and due to the increase in international oil price. Together with general lack of raw materials, the question of finding alternative concepts and technologies for energy production and consumption - energy-saving included - gained a high priority also in the HV A's activities. The party leadership increasingly expected the HV A to gather know-how and scientific-technical knowledge, an expectation the HV A sought to fulfill by intensifying its scientific-technical intelligence. [37]

These changes in the content structure of the report corpus are supported by structural changes over time in the set of keywords used for summarizing the reports, as evidenced by the over 60 percent decrease in the total number of keywords. The content also changed radically. The set of keywords used between 1980 and 1984 contained 63 percent of the keywords used between 1975 and 1979. The set of keywords used from 1985 onwards contained 45 percent of the keywords from 1975-1979, and only 39 percent of those used in the first half of the 1980s. However, the core keywords - *CSCE*, *foreign policy*, *domestic policy* - remain steady over time and can thus be regarded as the most important thematic path in the report corpus.

The keyword co-occurrence network visualizations for the most part support the results from the content structure analysis (Figure 4). There are, however, some observable differences between centrality metrics and positions in the graph worthy of discussion. In the graph showing the keyword co-occurrence network of 1975-1979, the keywords *communist party* and *evaluation* (label EINSCHAETZUNG) ranked as first and third according to betweenness centrality metrics and are pushed as hubs to the graph periphery (see Figure 4, sub-figure (a)). The graphs for 1980-1984 and 1985-1989 demonstrate a rather scattered co-occurrence structure, though they confirm the importance of keywords with the highest betweenness centrality metrics (see Figure 4, sub-figure (b) and (c)). This seems to be quite logical, since several keywords are included in approximately the same amount of reports, thus making it difficult for the layout algorithm to find clear authorities.

What is actually the story the network graphs visualizing the dynamics of the report corpus are telling about the historical period of 1975-1989? The changes detected in the thematic structure of the reports can be read as a map of challenges the GDR was facing. In the second half of the 1970s, the CSCE clearly dominated the political agenda of the GDR. From the early 1980s onwards, economic issues and questions related to international politics became more significant. The developments in the second half of the 1980s clearly illustrate how East German foreign intelligence became more and more harnessed for finding solutions to domestic problems of the GDR, mostly in the technological domain. Thus, the reporting of the HV A revolved to a greater extent around economic and energy-related issues, but also around West European integration. However, taking into account the limited number of reports in each period, the network graphs seem to suggest a shift away from thematically focused reports in the direction of general assessments of the current situation dealing with a wide range of questions. The visualizations also show that in regard to Nordic affairs the HV A was increasingly engaged in gathering information and compiling reports and analyses on economic problems and their solutions. Since these problems are known to have played an important role in the demise of communism[38], the HV A seems to have recognized the danger, yet not been capable of producing any usable solutions.[39]

36 E.g. Schroeder, 1998, 233ff.; Gieseke, 2008
37 See Gieseke, 2001, 210-214
38 E.g. Wallander, 2003; Gaddis, 2005; Rafalzik, 2010
39 See also Müller-Enbergs, 2010, 111

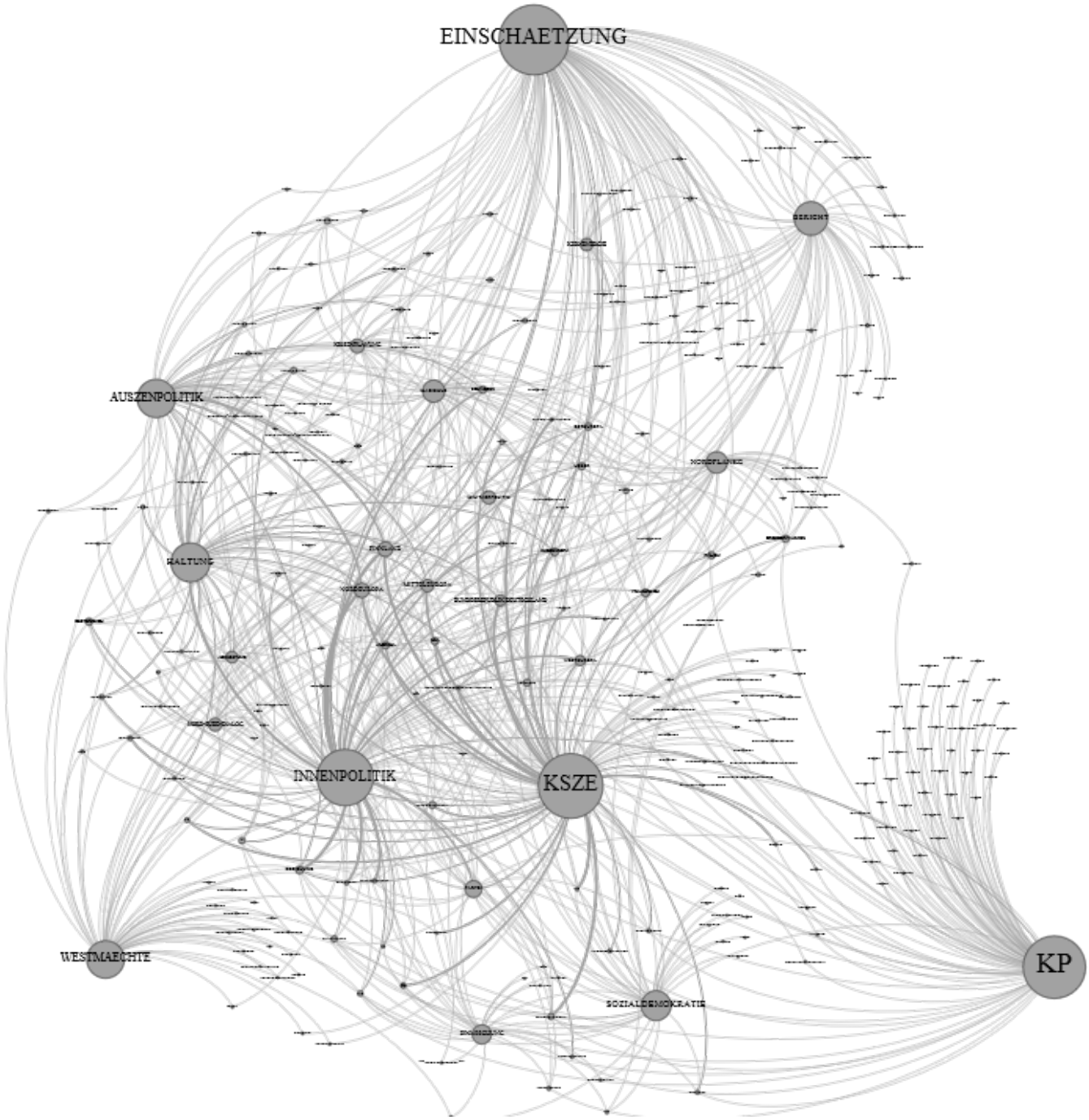


Figure 4A: 1975-1979

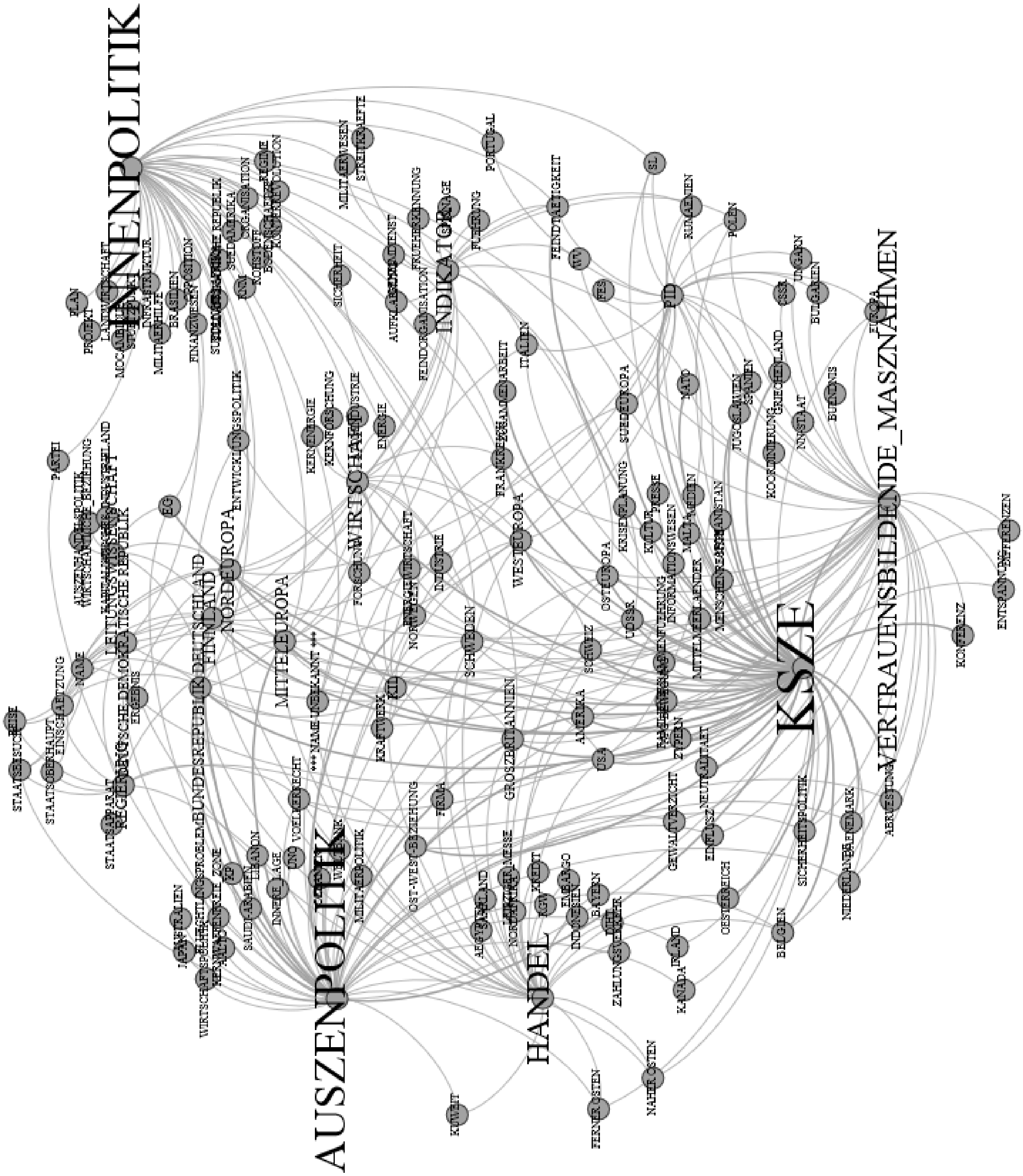


Figure 4B: 1980-1984

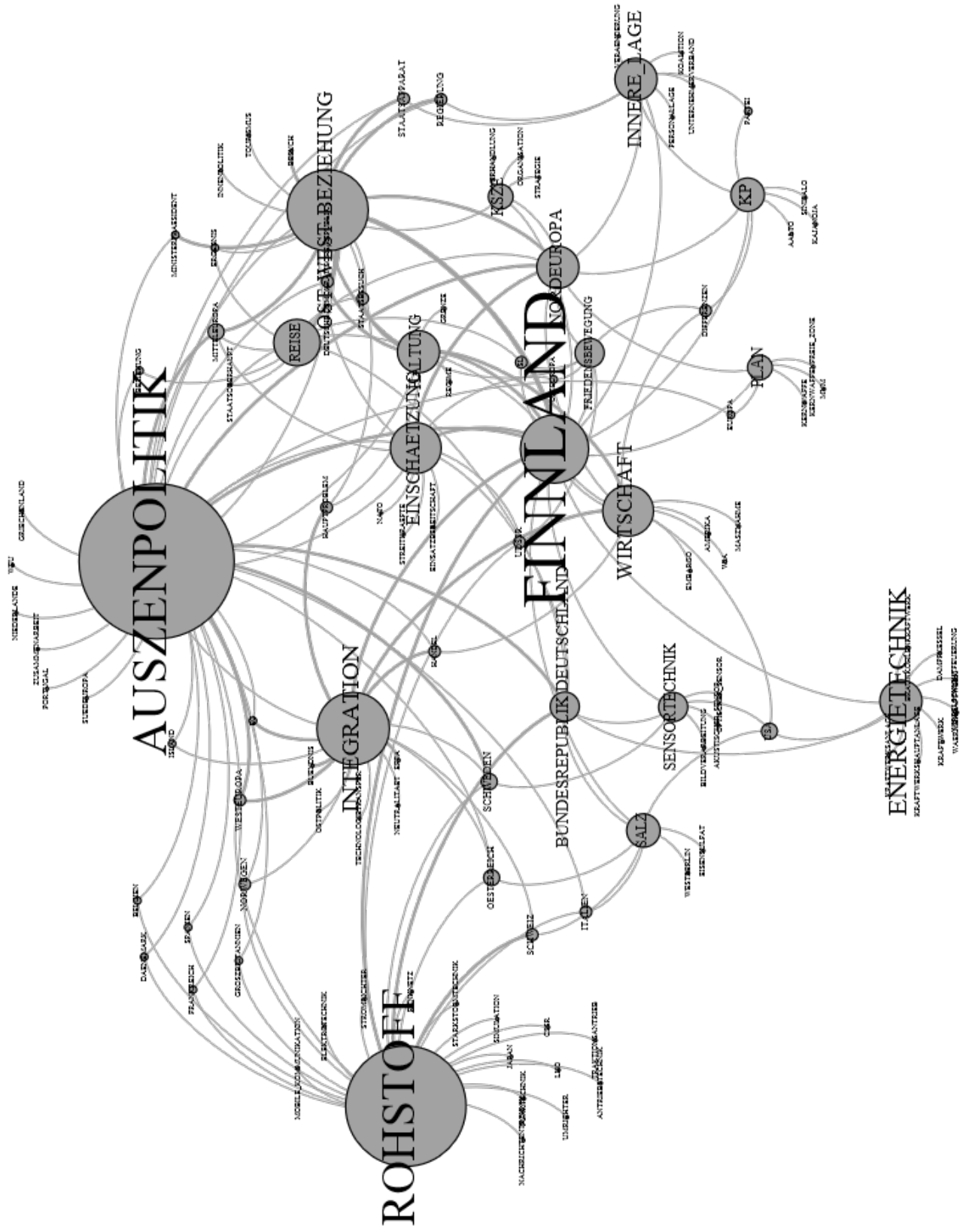


Figure 4C: 1985-1989

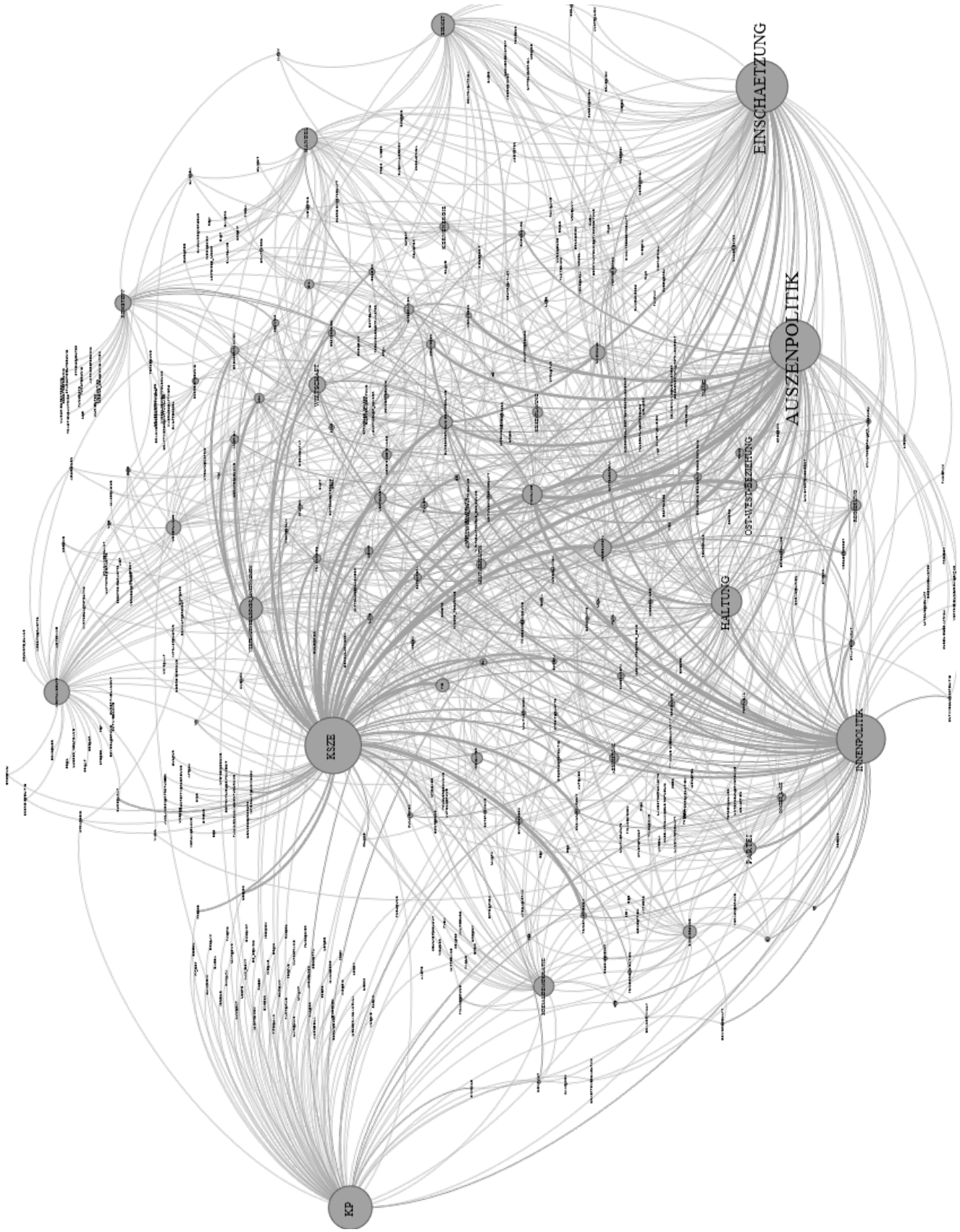


Figure 4D: Overall (1975-1989)

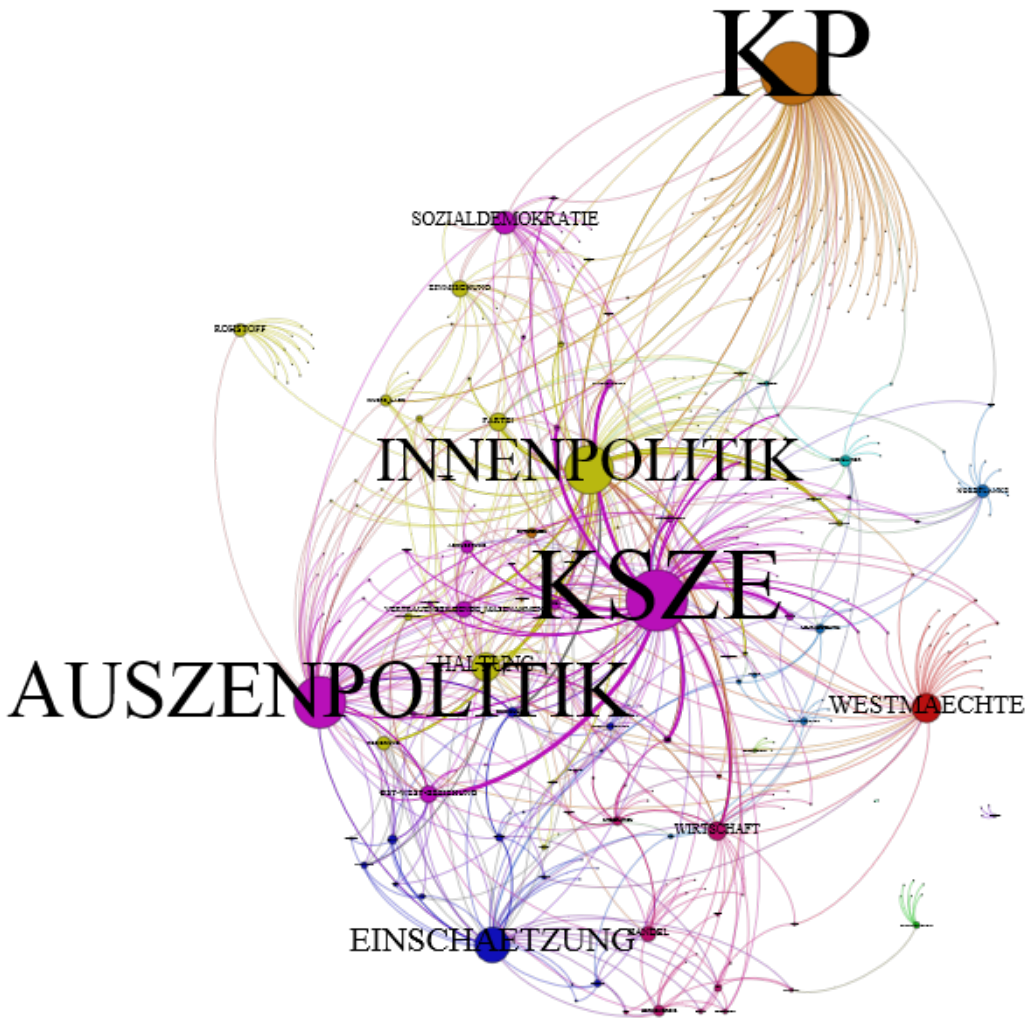


Figure 5A: Modularized report corpus network 1975-1989.

The issue of network clusters has already been touched upon above, as both the network visualizations and network metrics indicated the existence of different thematic clusters in the intelligence report corpus. In general, the growing availability of digitalized historical materials and on-line resources, data classification and categorization (including community, clique or sub-network detection) as methods to focus the research on critical points and structures have gained in importance. However, despite extensive studies[40], community detection has remained a core problem and theoretical challenge in network analysis. This article will utilize the community detection mechanism based on modularity also known as ”Louvain method”, which relies on the assumption that nodes being more densely connected to each other than with the rest of the network construct a cluster or community.[41] Since Gephi has a built-in support for this method, its application is quite straightforward.[42]

Community network metric imply the existence of three main content clusters. The largest cluster is comprised of 21.65 % of the total keywords and is built around the keywords *CSCE*, *foreign policy* and *social democracy*. The second and third content clusters include 20.96 % and 18.56 % of the total keywords, respectively. The second largest content cluster is characterized by the keywords *domestic policy*, *raw materials* and *standpoint*. As for the third content cluster, this cluster revolves around the keywords *communist party*, *differences* and *peace movement*. It is worth noting that two keywords - *CSCE* and *communist party* – which belong to different clusters have the highest degree, and thus, seem to have a stronger influence on the co-occurrence structure than other keywords. The two clusters built around these two keywords have different semantic structures, and

40 For excellent summaries of recent discussions, see Fortunato, 2010; Wu et al., 2013
41 See Blondel et al., 2008
42 https://sites.google.com/site/findcommunities/ [on-line. Last visited on 9th October, 2014].
43 Newman, 2006; Paranyushkin, 2011, 18. Cf. Du et al., 2008

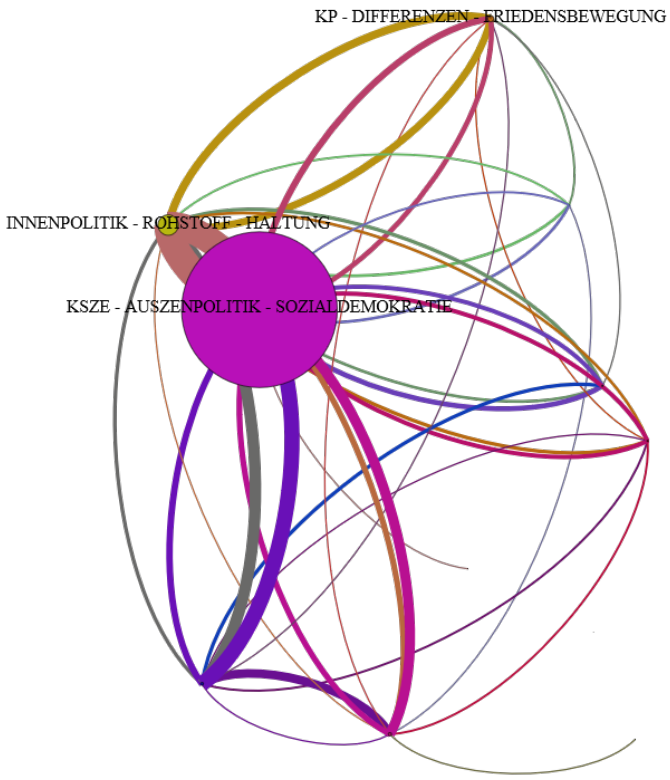


Figure 5B: Modularized and grouped report corpus network (1975-1989).

consequently, differ also in content. The cluster revolving around *CSCE* seems to have its focus on Western influence and international politics whereas the cluster built around *communist party* seems to deal with threats and challenges within the Communist camp.

The graph above visualizes the modularized community structure of the keyword co-occurrence network in two different forms, where each community has its own color (see Figure 5). The sub-figure (a) is an ungrouped visualization, i.e. the keywords are colorized but not grouped. Already this visualization implies the existence of several thematic clusters in the report corpus. The content cluster structure becomes more clear in the second visualization based on grouped keywords (Figure 5, sub-figure (b)).

4. Concluding remarks

This article sought to exemplify how social network analysis could be applied to keyword co-occurrence networks in order to uncloak hidden content structures in a corpus of intelligence reports and to understand continuity and change within the corpus. The methodology was based on two assumptions. First, all kinds of texts can be presented as networks capable of capturing the most important content. Second, text networks can be analyzed as social networks of language in which relationships between concepts or words are analyzed in a similar manner as relationships between individuals in social networks. From this perspective, concept co-occurrence networks not only embody the most relevant information, but also allow comparisons in time, and thus, help to focus on continuity and change. As the analysis presented in this paper shows, SNA offers powerful tools for identifying the most influential (central) keywords that function as junctions in the content structure, for detecting distinct communities present in the report corpus, and for comparing networks in order to examine change. Further, comparisons with prior historical knowledge illustrated the reliability of the results produced by the new methodology. The latter is especially important in historical network research. A critical contextual assessment- ”Does this make sense in this context?” - can protect from overestimating but also from underestimating the results. One should keep in mind that historical network analysis change the way in which a historian looks at the material, but also affects what they see. Therefore, results challenging previous knowledge can be exactly that - new knowledge.

The results underline the role of visualizations in the process of producing and presenting new knowledge. On the one hand, graphs are an important tool for validating the results of the network metric analysis and for focusing the future research. Graphs can also be used for exploratory analysis; instead of using networks analysis to answer ready-formulated research questions or for hypothesis testing, one could also ”let the graphs speak” in order to explore new views on old data or to generate new hypotheses. On the other hand, visualizations also verified the important impact of visualization layouts for knowledge production. Layouts should be understood as different views on the same network, views produced by different algorithms and parameters. Unfortunately, in many publications, visualizations are rarely used for presenting the same network from different perspective, but just for graphically presenting the network. Since layouts alter our perception of the network, a proper understanding of the visualization algorithm is indispensable for a critical assessment of the visualized knowledge. A proper layout selection can result in new knowledge while an unsuitable layout can nullify the results.

Two crucial aspects might limit the use of computational methods and algorithms among historians (or humanists in general). The first one is related to education and research training; humanists are less familiar with computational science than their colleagues

e.g. in the social sciences, and consequently, often more skeptical about the benefits of Digital Humanities. In this sense, this paper should be read as an encouragement for Digital Humanities inspiring colleagues to leave the ”comfort zone” and to familiarize themselves with methods and tools available for computational research in the humanities. The second aspect is related to the question of suitable sources for network data. To be sure, data for small-scale networks can be entered manually. However, the analytical step towards network analysis of large-scale, complex networks can be taken only through digitalization and computational processing of historical sources, which is technically demanding, costly and time-consuming. Hopefully, this article has shown that Digital Humanities may be the multi-disciplinary window to a new analytical landscape, to a new understanding of historical events - not as a substitute for, but as a complementary to existing research techniques and methodologies.

References

Alter, Steven. (2008). Defining Information Systems as Work Systems: Implications for the IS Field. *European Journal of Information Systems*, 17(5), 448–469.

Avison, David E., & Myers, Michael D. (1995). Information Systems and Anthropology: an Anthropological Perspective on IT and Organizational Culture. Information Technology & People, 8(3), 43–56.

Bastian, Mathieu, Heymann, Sebastien, & Jacomy, Mathieu. (2009). *Gephi: An Open Source Software for Exploring and Manipulating Networks*. Online: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154/1009

Blondel, Vincent D., Guillaume, Jean-Loup, Lambiotte, Renaud, & Lefebvre, Etienne. (2008). Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*. Online: http://arxiv.org/pdf/0803.0476v2

Brier, Alan, & Hopp, Bruno. (2011). Computer Assisted Text analysis in the Social Sciences. *Quality & Quantity*, 45(1), 103–128.

Brown, Archie. (1996). *The Gorbachev Factor*. Oxford: Oxford University Press.

Bruce, James B., & George, Roger Z. (2008). Intelligence Analysis - The Emergence of a Discipline. In: George, Roger Z., & Bruce, James B. (eds), *Analyzing Intelligence: Origins, Obstacles, and Innovations*. Washington, DC: Georgetown University Press, 1-15.

Diesner, Jana, Carley, Kathleen M., & Tambayong, Laurent. (2012). Extracting socio-cultural Networks of the Sudan from Open-source, Large-scale Text Data. *Computational and Mathematical Organization Theory*, 12(3), 328–339.

Du, Haifeng, White, Douglas R., Ren, Yike, & Li, Shuzhuo. (2008). *A Normalized and a Hybrid Modularity*. Online: http://intersci.ss.uci.edu/wiki/pub/20080401drwNormalizedModularityDraft_BW.pdf

Eklund, Tomas, Toivonen, Jarmo, Vanharanta, Hannu, & Back, Barbro. (2011). *Customer Feedback Analysis Using Collocations*. Amcis 2011 Proceedings - All Submissions, Paper Nr. 158.

Enders, Walter, & Su, Xuejuan. (2007). Rational Terrorists and Optimal Network Structure. *Journal of Conflict Resolution*, 51(1), 33–57.

Feicheng, Ma, & Yating, Li. (2014). Utilising Social Network Analysis to Study the Characteristics and Functions of the Co-Occurrence Network of Online Tags. *Online Information Review*, 38(2), 232–247.

Fortunato, Santo. (2010). Community detection in graphs. *Physics Reports* 486, 75–174. Online: http://arxiv.org/pdf/0906.0612v2

Friis, Thomas Wegener, Macrakis, Kristie, & Müller-Enbergs, Helmut (eds). (2010). *East German Foreign Intelligence: Myth, reality and controversy*. London, New York: Routledge.

Gaddis, John Lewis. (2005). *The Cold War. A New History*. New York: The Penguin Press.

Garthoff, Raymond L. (2004). Foreign Intelligence and the Historiography of the Cold War. *Journal of Cold War Studies*, 6(2), 21–56.

Gieseke, Jens. (2001). *Mielke-Konzern: Die Geschichte der Stasi 1945-1990*. Stuttgart/München: dva.

Gieseke, Jens. (2008). East German Espionage in the Era of Détente. *Journal of Strategic Studies*, 31(3), 395–424.

Herman, Michael. (2001). *Intelligence Services in the Information Age*. London: Frank Cass.

Holt, Tracy Van, Johnson, Jeffrey C., Brinkley, James D., Carley, Kathleen M., & Caspersen, Janna. (2012). Structure of Ethnic Violence in Sudan: a Semi-automated Network Analysis of Online News (2003-2010). *Computational and Mathematical Organization Theory*, 12(3), 340–355.

Hsu, Yi-Yu, & Kao, Hung-Yu. (2013). CoIN: a Network Analysis for Document Triage. Database, 1–11.

Hutchins, ChristopherE., & Benham-Hutchins, Marge. (2010). Hiding in Plain Sight: Criminal Network Analysis. *Computational and Mathematical*

Organization Theory, 16(1), 89–111.

Hyvönen, Timo, Järvinen, Janne, & Pellinen, Jukka. (2008). A Virtual Integration - The Management Control System in a Multinational Enterprise. *Management Accounting Research*, 19(1), 45–61.

Konopatzky, Stephan. (2003). Möglichkeiten und Grenzen der SIRA-Datenbank. In: Herbstritt, Georg, & Müller-Enbergs, Helmut (eds), *Das Gesicht dem Westen zu. DDR-Spionage gegen die Bundesrepublik Deutschland*. Bremen: Edition Temmen, 112–132.

Krebs, Valdis E. (2002). Uncloaking Terrorist Networks. *First Monday*, 7(4). Online: http://firstmonday.org/ojs/index.php/fm/article/view/941/863

Lee, Sangno, Song, Jaeki, & Kim, Yongjin. (2010). An Empirical Comparison of Four Text Mining Methods. *Journal of Computer Information Systems*, 51(1), 1–10.

Malm, Aili, & Bichler, Gisela. (2011). Networks of Collaborating Criminals: Assessing the Structural Vulnerability of Drug Markets. *Journal of Research in Crime and Delinquency*, 48(2), 271–297.

Marres, Noortje. (2012). The Redistribution of Methods: On Intervention in Digital Social Research, broadly Conceived. *The Sociological Review*, 60(S1), 139–165.

Morselli, Carlo. (2010). Assessing Vulnerable and Strategic Positions in a Criminal Network. *Journal of Contemporary Criminal Justice*, 26(4), 382–392.

Musiał, Kazimierz. (2009). Reconstructing Nordic Significance in Europe on the Threshold of the 21st Century. *Scandinavian Journal of History,* 34(3), 286–306.

Müller-Enbergs, Helmut. (1998). *Inoffizielle Mitarbeiter des Ministeriums für Staatssicherheit, Teil 2: Anleitungen für die Arbeit mit Agenten, Kundschaftern und Spionen in der Bundesrepublik Deutschland*. Berlin: Links, 2nd edition.

Müller-Enbergs, Helmut. (2007). *"Rosenholz" Eine Quellenkritik*. Die Bundesbeauftragte fur die Unterlagen des Staatssicherheitsdienstes der ehemaligen Deutschen Demokratischen Republik, Berlin (BF Informiert 28).

Müller-Enbergs, Helmut. (2008). *Die Inoffiziellen Mitarbeiter (MfS-Handbuch)*. Berlin: BStU.

Müller-Enbergs, Helmut. (2010). Political Intelligence: Foci and Sources, 1969-1989. In: Friis, Thomas Wegener, Macrakis, Kristie, & Müller-Enbergs, Helmut (eds), *East German Foreign Intelligence: Myth, reality and controversy*. London, New York: Routledge, 91–112.

Müller-Enbergs, Helmut. (2011). H*auptverwaltung A (HV A). Aufgaben – Strukturen – Quellen (MfS-Handbuch)*. Berlin: BStU.

Newman, Mark E. J. (2006). Modularity and Community Structure in Networks. *Proceedings of the National Academy of Sciences*, 103(23), 8577–8582.

Noorbehbahani, F., & Kardan, A. A. (2011). The Automatic Assessment of Free Text Answers Using a Modified Bleu Algorithm. *Computers & Education*, 56(2), 337–345.

Novotny, Josef, & Cheshire, James A. (2012). The Surname Space of the Czech Republic: Examining Population Structure by Network Analysis of Spatial Co-occurrence of Surnames. *Plos one*, 7(10), e48568.

Oesper, Layla, Merico, Daniele, Isserlin, Ruth, & Bader, Gary D. (2011). *Wordcloud: a Cytoscape Plugin to Create a Visual Semantic Summary of Networks*. Source code for Biology and Medicine, 6(7).

Özgür, Arzucan, Cetin, Burak, & Bingol, Haluk. (2008). Co-occurrence Network of Reuters News. *International Journal of Modern Physics*, 19(5), 689–702.

Paranyushkin, Dmitry. (2011). *Identifying the Pathways for Meaning Circulation using Text Network Analysis*. Research report. Nodus Labs. Online: http://noduslabs.com/publications/Pathways-Meaning-Text-Network-Analysis.pdf

Prell, Christina. (2012). *Social Network Analysis*: *History, theory and methodology*. London: SAGE.

Raab, Jörg, & Milward, H. Brinton. (2003). Dark Networks as Problems. *Journal of Public Administration Research and Theory*, 13(4), 413–439.

Rafalzik, Sascha. (2010). *Wirtschaftsspionage* der DDR. Münster: LIT.

Schroeder, Klaus. (1998). *Der SED-Staat. Geschichte und Strukturen der DDR*. München: Bayerische Landeszentrale für politische Bildungsarbeit.

Schultz-Jones, Barbara. (2009). Examining Information Behavior Through Social Networks: An Interdisciplinary Review. *Journal of Documentation*, 66(4), 592–631.

Schwartz, Daniel, & Rouselle, Tony. (2009). Using Social Network Analysis to Target Criminal Networks. *Trends in Organized Crime*, 12(2), 188–207.

Scott, John P. (2013). *Social Network Analysis*. London: Sage Publishing, 3rd edition.

Steinbock, Dan. (2008). NATO and Northern Europe: From Nordic Balance to Northern Balance.

American Foreign Policy Interests, 30(4), 196–210.

Stuart, Keith, & Botella, Ana. (2009). Corpus Linguistics, Network Analysis and Co-occurrence Matrices. *International Journal of English Studies*, 1–20.

Wallander, Celeste A. (2003). Western Policy and the Demise of the Soviet Union. *Journal of Cold War Studies*, 5(4), 137–177.

Walsh, Patrick F. (2011). *Intelligence and Intelligence Analysis*. Routledge, New York. Abingdon: Routledge.

Wu, Zhiang, Cao, Jie, Wu, Junjie, Wang, Youquan, & Liu, Chunyang. (2013). Detecting Genuine Communities from Large-Scale Social Networks: A Pattern-Based Method. *The Computer Journal,* 1–15.

Xu, Jennifer, & Chen, Hsinchun. (2005). Criminal network analysis and visualization. *Communications of the ACM,* 48(6), 100–107.

Yang, Libin, Cai, Xiaoyan, Zhang, Yang, & Shi, Peng. (2014). Enhancing Sentence-level Clustering with Ranking-based Clustering Framework for Theme-based Summarization. *Information sciences*, 20, 37–50.

DEN
Data Exchange Network

The Danish Elite Network

Christoph Houman Ellersgaard
University of Copenhagen
Copenhagen, Denmark

Anton Grau Larsen
University of Copenhagen
Copenhagen, Denmark

Abstract
This article presents the extensive Danish elite network. Collected during 2012 and 2013, the data comprises 56,536 positions within 5,079 affiliations, and connects 37,750 individuals. The network consists of the largest Danish corporations, state institutions, NGO's, and other integrative networks such as social clubs or royal events. Data were gathered through an inclusion principle, adding all potentially interesting affiliations. Procedures of name-matching and quality control are presented. Finally, the data are introduced: made available through a package for R, which enables the creation of subnetworks and weights.

Keywords: Elite networks, Corporate interlocks, Policy networks

Authors
Christoph Houman Ellersgaard, is a PhD Fellow at the Department of Sociology, University of Copenhagen. His research interests include economic elites, social stratification and methodological issues of power elite identification using social network analysis and multiple correspondence analysis.

Anton Grau Larsen, is a PhD Fellow at the Department of Sociology, University of Copenhagen. In his dissertation, Anton Grau Larsen explores new methods and data sources with focus on cohesion between elite groups and strategies for handling extensive network data on power networks though statistical programming in R.

Correspondence concerning this work should be addressed to: Christoph Houman Ellersgaard, Department of Sociology, University of Copenhagen, Østre Farimagsgade 5, Building 16.2.42, 1014 Copenhagen K, Denmark. Email: che@soc.ku.dk

Acknowledgements
Thanks to Business Information firm BIQ for providing data on the largest Danish corporations and foundations; to journalist Annelise Weimann for providing access to her archive of Danish royal events; and to Hanne Dahl and Gads Forlag, publishers of the Danish equivalent of Who's Who Kraks Blå Bog, for providing access to the biographical material on the core of the elite network. We are also grateful to IT-specialist Martin Kjeldgaard Nikolajsen, Department of Sociology, University of Copenhagen, and Jacob Aagaard Lunding for assisting us during the data collection.

DOI: http://dx.doi.org/10.17266/35.1.5

1. Introduction

Since C. W. Mills wrote his seminal work, The Power Elite, elite research has struggled to define elites empirically. In 1956, Mills offered a compelling definition of the power elite as: '*…those political, economic and military circles which as an intricate set of overlapping cliques share decisions having at least national consequences*' (Mills 1956:18). Mills himself had to rely on the positional method (see Knoke, 1993) and could not map the overlapping circles directly.

By collecting a two-mode network of all nationally integrating publicly available and official affiliations in the small nation-state of Denmark, 'The Danish Elite Network' dataset permits the identification of the central individuals, affiliations and cross-cutting social circles that compose the power elite. (See Ellersgaard and Larsen 2014; 2015). The data can easily be split into parts according to sectors such as politics, business, organisations and state, allowing more detailed studies of sub-elites, but still with reference to the total structure. Because Denmark is a small society with strong traditions of transparency in decision-making processes and widespread use of websites for both public and private organisations, it was possible to attempt a complete registration of all official elite affiliations. Collected during 2012 and 2013, the data contain 56,536 positions in 5,079 affiliations and connect 37,750 individuals. All nationally relevant official positions, such as company boards, committees, foundations, and advisory boards are included.

In the following sections we present the data collection process with its strengths and weaknesses, and we introduce the available data files. Finally, we present a table with the data details.

2. Data: The Extensive Elite Network

The data were collected in four fairly discrete phases:

1. The generation of lists of affiliations
2. Affiliation collection
3. Name-matching and quality control
4. Tagging

2.1 The generation of lists of affiliations

First, we created large initial lists of organisations and groups. The included groups are not affiliations themselves but lists of names of organisations and groups, their type, their addresses, and preferably their website. These initial lists provided the basis for a small snowball procedure for each organisation. Using self-reported organisational information, such as organisational

Table 1: Affiliation Networks Included in the Data

	From original sources	Excluded*	Added in snowball sample†	Final number of affiliation networks	No. of relations
State[1]	2,306	1,770	314	850	10,254
Parliament[1]	153	101	31	83	970
NGO[1]	749	181	922	1,490	16,434
Corporations[2]	1,136	85	40	1,091	7,476
Foundations[2]	1,380	69	83	1,394	8,181
VL Networks[3]	117	3	0	114	3,845
Commissions[4]	116	44	0	72	1,121
Events[5]	74	65	8	17	8253
Total	6,031	2,350	1,398	5,079	56,536

1 Source: Danish Public Administration Database (www.foa.dk)
2 Source: The list of largest corporations according to turnover, and the list of all foundations in the Danish Central Business Register (obtained through www.biq.dk)
3 Source: Homepage of VL networks (www.vl.dk)
4 Source: Registration of political commissions from 2005-2011, made by Danish weekly newsletter A4.
5 Sources: Webpage of Danish Royal Family (www.kongehuset.dk) and private archive of journalist.
*Affiliation networks were excluded if: no board of extra-organisational members existed, no information on the board was available (either online or through personal contact), the board was included in other sources (i.e. data was not duplicated), or the board entirely overlapped with another board within the same organisation.
† Includes both sub-committees within the organisations on the original list and the 142 networks obtained by snowballing the affiliations of prominent agents.

diagrams, we identified all the relevant committees, boards, sub-committees, board of representatives, and other groups. The names of all members were collected, along with their affiliation, role (e.g. chairman, director), a long string with their description (if available), and the link to the website describing of the affiliation. The initial lists of organisations and groups were constructed according to the inclusion principle, which dictated that all possibly relevant organisations were included in the dataset. This was ensured by constructing the lists from exhaustive official databases.

In Table 1, the organisations in the initial lists are grouped according to their source . Organisations were excluded if they: 1) did not have affiliations, like a board or a committee, 2) only had affiliation members from within the organisation, like coordinating groups for employees, 3) were overlapped entirely with another affiliation, either in the form of a duplicate entry or a subsidiary. The final number of affiliations from each source is the number remaining after removing those that formed no ties between organisations and adding the snowball samples. The list of state organisations was drawn from a large public database of governmental and non-governmental entities, the – Danish Public Administration Database (FOA, www.foa.dk). The FOA contains a hierarchical database placing each of the 'offices' in the Danish state organised in a nested structure, with names of leading personnel, addresses and sub-offices. All offices working at the regional level or above were made into an initial list. This excludes 'offices' working at the local level, such as public schools, individual churches, police and fire departments, but includes high schools and large hospitals. The state offices are often governed only by the office one step higher in the structure and as a result, many do not form boards or committees. This seems to be reserved for more autonomous entities higher in the structure.

The list of commissions was produced by the weekly newspaper A4. The list contains all commissions from 2005 until 2011, and was supplemented by the authors to 2013.

The list of parliamentary committees was taken from the official parliamentary website along with other institutions tied to the parliament.

Corporate boards were from the top 1,000 corporations according to turnover. Furthermore, the boards of independent subsidiaries of the top 1,000 corporations were added to the data. Structurally important corporations such as media and financial institutions were then added to the list. Both corporations and a complete list of boards of foundation originated from The Central Business Register (CVR). Advisory boards and sub-committees of the major foundations were added to the list of foundation boards.

The list of NGOs was drawn from the FOA database and includes all organisations with the right, given by the state, to be consulted on legislation. All unions, employers organisations, national sports associations, environmental groups, animal rights groups, and many more are included in this list. To the NGO list were added various organisations that did not fit with other lists, such as publicly known elite networks. The largest of these, the VL groups, is split into a separate list. The list of events includes all publicly listed balls and official dinners, and royal hunting parties held by the royal family from 2009 to 2013. The collected events differ considerably in size from the rest of the affiliations. The state affiliations have on average 11 members, whereas the events have around 200. Furthermore, three events have more than 1,000 participants.

2.2 The inclusion principle: Connecting Organisations at a National Level

Three rules of inclusion guided the many small snowball samples:

1. All affiliations should be able to connect individuals across organisations. This excludes affiliations that are reserved to employees of an organisation or that are purely internal, see above.
2. All affiliations must meet physically at least once a year and therefore create face-to-face interaction between their members .
3. All affiliations must operate at a regional or national level. They operate at a regional or national level if they integrate individuals at a regional or national level, thereby excluding local affiliations such as environmental groups working for the preservation of a particular forest or a local shelter for the homeless.

All three rules can be ambiguous in their application, but this is resolved by the general inclusion principle: if two data collectors disagree as to whether an affiliation should be included in the data, it is included.

2.3 Name-matching and quality control

Most individuals, 78%, in the dataset are only members of a single affiliation. The names of those with more than one membership were matched and confirmed from more than one source. These sources would often be

The relations matrix has the following variables:

NAME	The matched name of the individual. All names are unique.
AFFILIATION	The name of the affiliation with its function in parenthesis.
ROLE	The role of the individual within the affiliation. This information is not always included for the top 1,000 corporations.
GENDER	The gender of the individual. The gender is determined by the first name.
DESCRIPTION	The description of the individual taken from the affiliation website in relation to this position.
SOURCE	The source of each case and affiliation.
BIQ_LINK	If available it is the link to the BIQ.dk database. BIQ is a proprietary database with all present and past members of all corporations in Denmark along with annual reports.
CVR	The CVR number for each corporation and foundation. All Danish corporations have a CVR number. With the CVR number it is possible to match and collect a large variety of publicly available data.
TAG1-7	Each affiliation is coded with between one and seven of 278 thematic tags.

4. Data Details

Response rate	N/A
Non-respondent bias	N/A
Theoretical grouping	No questionnaire was used
Publications using these data	Stahl and Henriksen (2014) Ellersgaard and Larsen (2013, 2014, 2015)
Data context	Database of a national elite
Respondents	
Longitudinal	No
Temporality	Most positions last several years, although individual careers end within days. Some data points are separated by up to 2 years, for events up to 4 years and commissions up to 8 years
Analytical or pedagogical utility	The data allow for analysis of inter- or cross-sectorial ties. As they are divided into many sectors, students can choose sectors that interest them.
Known issues	Underestimates the amount of connections due to the name-matching procedure.

the collected descriptions from the affiliation websites, but could also be addresses and official registers. But this procedure slightly underestimates the ties, erring on the side of caution. The name-matching process was performed by sorting the list according to first name, last name, and full name. When two people hold the same name their names were given a numerical suffix, like "Hans Jensen 1" and "Hans Jensen 2".

In some instances, an individual may not use the same name in all affiliations; most commonly, a middle name was not used. The sorting procedure captures this practice but is very vulnerable to people changing their first name, such as using a middle name as a first name. If this practice was suspected, a search for possible alternative first names was made; however, it is impossible to achieve perfect-name-matching quality and the data therefore underestimates the number of connections.

2.4 Tagging: Ordering data

All affiliations in the data were tagged with up to seven tags. The tags are thematic: culture, music, science, education, social politics, and foreign relations. Tagging is different from categorising, because tags are non-exclusive and the number of tags for each affiliation varies. The tags were based on affiliation descriptions from the web-pages. All tags were controlled for consistency by two coders. The network can be split into sector by combining the relevant

tags into subjects and then extracting all affiliations with a relevant tag. The affiliations related to the Danish state can, for instance, be extracted with this collection of tags: 'State administration', 'Ministry', 'State corporation', 'Military', 'Public leaders'.

3. Data Files and Formats

Data is provided in the Danish_Elite_2013_Relations.csv file separated with "|" and encoded in UTF-8 and as an excel file. It is organised as a case-affiliation edge list with attributes attached to each case.

The dataset is also made available via the R package soc.elite, currently available on Github: github.com/antongrau/soc.elite. The package is based on igraph and includes functions for sub-setting by tags, descriptive functions, and functions for cleaning, coding and plotting. Handling the particular structure of the tags is made considerably easier by the function in the package, such as has.tags and tag.network.

In the package there are datasets that can be merged with 'The Danish Elite Network', such as biographical data on a core of 423 individuals. A dataset with biographical data on a core of 171 core business leaders and a dataset with data on size, turnover, number of employees and the like, on the top 1,042 corporations. The package also serves as a valuable tool in courses on social network analysis or elite sociology. Researchers who analyse the data available in soc.elite are encouraged to send their code to the package maintainer and it will be published in the package along with proper citations.

5. References

Ellersgaard, C. H., & Larsen, A. G. (2013). The inner circle revisited – the case of egalitarian society. Presented at the XXXIII Sunbelt Social Networks Conference of the International Network for Social Network Analysis (INSNA), Hamburg.

Ellersgaard, C. H., & Larsen, A. G. (2014). Identifying power elites – a social network analytic approach. Presented at the Understanding the transformation of economic Elites in Europe, Lausanne.

Ellersgaard, C. H., & Larsen, A. G. (2015). The Power Elite in the Welfare State – Key institutional orders of the power networks in Denmark. Working Paper. Department of Sociology, University of Copenhagen.

Knoke, D. (1993). Networks of elite structure and decision making. Sociological Methods & Research, 22(1), 23–45.

Mills, C. W. (1956). The Power Elite. Oxford: Oxford University Press.

Stahl, R. M., & Henriksen, L. F. (2014). Indlejret visdom: En netværksanalyse af det økonomiske råd. Politik, 17(2), 68–78.

Donor Motivations in the California State Legislature: A Social Network Analysis of Campaign Contributions

Jacob Apkarian
Virginia Tech
Blacksburg, VA USA

Robert Hanneman
University of California
Riverside, CA USA

Shaun Bowler
University of California
Riverside, CA USA

Byron Martin
Houston Community College
Houston, TX USA

Abstract
This paper investigates campaign donor motivations by examining the ways in which donations to political campaigns tie candidates for the legislature together. Using social network analysis, we examine campaign donations to the California state legislature in 2004. We demonstrate that donor motivations vary by donor type where individual donors and candidate committees are more ideologically motivated while business donors tend to be investment-driven when contributing. We also examine the extent to which different donor types reflect local interests along with how agendas regarding candidate demographic characteristics influence patterns of donation (i.e. who donors prefer to connect). We find that donors generally tend to promote polarization by connecting candidates of the same political party, but that experienced candidates are very well connected and tend to bridge the party divide.

Keywords: Campaign contributions, social networks, candidates, donors, motivations

Authors

Jacob Apkarian, Department of Sociology, Virginia Tech, Blacksburg, VA USA.

Shaun Bowler, Department of Political Science, University of California, Riverside, CA USA.

Robert Hanneman, Department of Sociology, University of California, Riverside, CA USA.

Byron Martin, Department of Political Science, Houston Community College, Houston, TX USA.

Correspondence concerning this work should be addressed to: Jacob Apkarian, Department of Sociology, Virginia Tech, Blacksburg, VA 24061 USA. Phone:(540) 231-8971.

DOI: http://dx.doi.org/10.17266/35.1.3

1. Introduction

Recently, work has begun to examine social embeddedness in legislative and electoral settings (Peoples 2010; Peoples and Gortari 2008). Some of this work has demonstrated the influence of extra-legislative actors in relation to candidates and political parties. The term 'extra-legislative actors' refers to actors who are neither voters, nor legislators, but nevertheless play an important role in electoral and even legislative politics (Koger, Masket, & Noel, 2009; Masket, 2007; Masket, 2009). In this paper we argue that the motives of key extra-legislative actors, specifically donors, can be indirectly observed via the network patterns of campaign contributions from donors to multiple candidates. Using social network analysis (SNA) we examine the extensive networks that tie candidates together via donations in the California state legislature in 2004. This approach treats donor behavior as linked across candidates rather than separated by candidate. Donors who support multiple candidates are a minority of donors, but supply most of the money in political campaigns. In supporting multiple candidates, such donors are often systematically pursuing an ideological, pragmatic, or expressive agenda.

We examine donation patterns by donor type (individuals, businesses, PACs, etc.) to explore donor motivations and how they influence polarization at the state level. We also explore the extent to which donor interests are local in scope along with how candidate demographics influence donation patterns.

1.1 Extra-legislative Actors

It has become clear in the literature that extra-legislative actors are an important influence in politics. "Given how crucial events outside the chamber are to behavior within it, it is remarkable that more attention is not devoted to such extra-legislative forces" (Masket, 2007, 483). These external actors "wield power over party nominations and the resources needed to win them" (Masket, 2007, p. 484) and hence can shape legislative behavior (Coffey, 2007; Cohen et al., 2008; Koger et al 2009; Gordon, 2001; Peoples, 2008; Snyder, 1990; Stratmann, 2002). In identifying how actors outside the legislature may seek to influence legislators, this literature implicitly provides an argument that legislators may be connected to each other via these extra-legislative actors. Legislators are likely embedded within a network in which they have extra-legislative actors in common. Knowing that voting patterns among legislators are tied to relations between them (Peoples, 2008), this fact is important due to the implications for policy decision making (Peoples & Gortari, 2008).

Levendusky notes that the argument identifying the importance of extra-legislative actors is both novel and appealing (Levendusky, 2009, p. 833). Part of the argument of Masket and others, for example, is that these extra-legislative actors help support and even promote polarization,but there is a need for greater specificity. Different actors may have diverse motivations that push legislators in different ways: some may push towards greater polarization while others may push away from it. Nor is it always clear how extra-legislative actors may be identified a priori. In other words, despite the appeal of the argument, it can sometimes be hard to see specific, concrete, and measurable examples of behavior exhibited by identifiable extra-legislative actors.

One way in which we can develop greater specificity to this argument is by reference to a concrete and measurable resource controlled by extra-legislative actors – that of money. Money is a valuable resource for any campaign (e.g. Brown, Powell, & Wilcox, 1995; Herrnson, 2008; Jacobson, 1980).[1] Campaign donations offer a good example of the broader argument of extra-legislative actors because they provide a concrete metric by which we can measure the behavior of those actors and examine relations between them. For our purposes, donations provide the metric by which we can identify ties within networks (for representative examples of work in this area that also include comprehensive reviews of the literature see Brown et al., 1995; Francia et al., 2003; Gimpel, Lee, & Kiminski, 2006, 2008; Hernnson, 2008; Jacobson, 1980; Malbin, 2003; Panagopolous & Bergan, 2006; Peoples & Gortari, 2008; see also websites such as the National Institute on Money in State Politics and the Campaign Finance Institute: www.cfinst.org).

Discussions of donations typically take an individual level approach to analyzing the decisions of donors without taking the behavior of other actors into consideration (Otte & Rousseau, 2002, p. 441). That is, a donation between a given donor and candidate X is often (albeit implicitly) seen to be independent from the relationship between the donor and candidate Y.[2] The argument concerning extra-legislative actors suggests that they act in concert and/or push different candidates in similar ways (Burris, 2001; Clawson, Neustadtl, &

1 Assembly Speaker Jess Unruh's comment that "money is the mother's milk of politics" is one of the more familiar quotations in California politics.
2 Similarly, donations to candidate X may be assumed, again often implicitly, to come from a different set of donors to those attracted by candidate Y.
3 E.g. "candidates must wage a campaign to raise money that is just as complicated as the campaigns they wage to win votes" (Francia et al., 2003, p. 69; see also Brown et al., 1995, p. 19-20).

Bearden, 1986). Therefore it makes sense to examine 2-mode candidate-donor networks where candidates are indirectly tied via shared donors.

The connections producing these candidate-donor networks are in part due to the practicalities of raising money for campaigns. The techniques candidates use to overcome difficulties in raising money[3] challenge the assumption that a given donation is likely to be independent from other donations. Campaigns use mailing lists and Rolodexes and not simply a telephone directory. Those who have given money in the past or have donated to another candidate in the current election would be more promising leads than a random name on a voter list. Donor names may show up in several different sources and their identity becomes known to (and hence become targets for) several candidates. Attendees of fundraising events are likely to run into donors they have seen at other similar events. Contributors, too, may consciously coordinate their actions across different legislators: As one PAC office admits "We talk to each other all the time. Are you giving to Henry or to Steve? Or you ought to be giving to Henry or Steve" (quoted in Peoples, 2010, p. 654).

Closer examination of these complex candidate-donor networks is needed. Some scholars have taken note of this kind of embeddedness but these have been largely informal characterizations. In their study of campaign finance, for instance, Brown et al. (1995) refer several times to networks (e.g. "personal acquaintance networks," p. 57) but that observation is not pursued more extensively.[4]

One way of describing the embeddedness of candidates in their relationships with campaign donors is to use the tools of social network analysis (SNA). SNA has attracted a great deal of attention of late in part because it can reveal the ways in which social actions – including those of legislators - may be embedded in broader contexts (e.g. Bowler & Hanneman, 2006; Fowler, 2006; Koger et al. 2009; Peoples, 2010; Peoples & Gortari, 2008; Siegel, 2009). SNA is particularly relevant in these situations where multiple actors may have relationships between them.[5] A SNA approach is useful for understanding the relationships between extra-legislative actors and candidates by providing a way for us to describe and characterize relationships using both statistical and graphical techniques.

Also, the network patterns identified speak to the relationship between extra-legislative actors and candidates allowing us to see, for example, how the networks of contributions may vary by candidate attributes and donor types. There is obviously variation in how much a donor may give, but the value of social network analysis is that it allows us to examine how different donor types give their money in different ways, therefore connecting candidates with certain attributes and not others. Most donors give small amounts of money to a single candidate, which is an individual expressive act. But many donors support multiple candidates, and are expressing a broader agenda and a deeper level of involvement. These multi-donors give more money, and are more influential. Who multi-donors support and who they do not, tells us something about their agendas and reasons for participation. By analyzing the dyads of candidates who are supported by specific types of donors, we can indirectly observe donor motivations.

1.2 Donor Types

We have identified six different types of donors who comprise extra-legislative actors in our study: individuals, candidate committees, social organizations, PACs, and two types of businesses (small and large). Individuals are individual citizens donating to a candidate. Candidate committees are voluntary organizations, usually managed by a candidate's agents or partisans (e.g. "Friends of X for Assembly"). Social organizations such as labor unions, Native American tribes, and occupational and industry associations generally aggregate the resources and interests of rather narrow classes of persons. Political Action Committees (PACs) have some similarities to social organizations in that they may represent fairly narrow classes (e.g. San Diego Dentist's Association PAC) but membership in these associations have a stronger voluntary component than for organizations such as unions.[6] Small businesses are generally controlled by individuals, partners, or a socially embedded group of owners organized as a small company.[7] Big businesses, however, are generally corporate in form, and often

4 More recent work (Gimpel et al., 2006; Gimpel et al., 2008; Cho & Gimpel, 2007) has begun to examine the political geography of campaign contributions and shows, among other patterns, that candidates from both parties raise money from similar geographic areas (e.g. Gimpel et al., 2006, p. 628). But this work has not examined the network attributes of links between candidates, nor the way in which these networks tie legislative and extra-legislative actors together.

5 In SNA it is the relationships between actors that become the main object of study: "Relational data are the focus of the investigations" (Otte & Roussea, 2002, p. 442; see also Knoke & Kuklinski, 1982; Hanneman & Riddle, 2005).

6 It should be noted that the distinction between many social organization donors and political action committee donors is far from clean and clear. PACs vary widely in the scope of interests represented, and may not really be very voluntary for members of represented classes. Occupational and professional associations, particularly, may appear in donor lists as either social organizations, or as PACs that are run by, and closely held by, social organizations. If a donor reported itself as a PAC, it was coded as such.

have publicly traded ownership rights.[8] Our overarching expectation is that networks are likely to vary by donor type.

1.3 Donor Motivations

We know from existing literature that motivations for donation will vary. Following the literature on campaign donations we may distinguish between ideology-driven motivation and investment-driven[9] motivation (Clawson et al., 1986). The distinction is between those donors that contribute as a means to achieve some broader goal based on values that they prescribe to and those that contribute with the goal of directly benefitting individually. We view these motivations as poles on a semantic differential scale. In reality, donor motivations are likely to be a mixture of these ideal types that manifest in different ways among different donor types.

Donors motivated by ideology are likely to support candidates whose parties share their values. Therefore we would expect that ideology-driven donors would be highly connected to candidates of a single party in the candidate-donor network. Rather than pursuing an ideological or partisan agenda, investment-driven donors would likely make contributions across party lines, as they are likely to engage in hedging behavior. Therefore, we would expect investment driven-donors to tie together candidates of different parties, and possibly even direct opponents.

Donor motivations also determine whether donors connect incumbents or challengers. In their research of campaign contributions from the business sector, Clawson et al. (1986) find that certain businesses are driven by individual interest (vs. the interest of the industry as a whole) and attempt to buy access via campaign contributions to "powerful incumbents" (p. 798). They demonstrate that other businesses are ideologically driven and support business friendly challengers. Based on this prior research, we would expect ideology-driven donors to be biased in favor of connecting challengers whereas investor-driven donors should connect experienced incumbents.

Certain ideology-driven donors are likely to be motivated to contribute in accordance with the demographic traits of candidates. Women and non-white candidates are certainly minorities in U.S. politics (Conti, 2002). In fact, it is in part an assumption that non-male and non-white candidates face fundraising difficulties that underlies explicit attempts by groups such as Emily's List[10] or the Wish List[11] to try and help raise money for these disadvantaged candidates. Therefore, we might expect certain donor types to center some of their support for multiple candidates on demographic characteristics like ethnicity or gender. Bratton and Haynie (1999) demonstrate that women and racial minorities "pursue distinctive legislative policies" (p. 672). Therefore donors that share similar political values and concerns might be more likely to be tied to multiple women or minority candidates. For example, research demonstrates that female politicians are more likely to promote legislation dealing with issues of gender equity as it relates to health care, poverty, and education, than their male counterparts (Bratton & Haynie, 1999; Lawless, 2004). Donors interested in advancing these types of policies would be expected to support multiple female candidates. Network analysis of contribution patterns can show whether or not fund-raising for women and minority candidates is distinctive from male and white candidates not simply with respect to amounts raised but the ways in which money may be raised.

There are other motivations for contributing that don't necessarily fit neatly onto either pole of ideology or investment. The scope of campaign contributions likely varies by donor type which reflects different donor agendas. Certain donors are more concerned with local conditions than others and are therefore likely to contribute to and thereby connect multiple candidates in the same geographic region for both ideological and instrumental reasons. We would expect donors with localized interests to connect candidates with overlapping constituencies, and have bias against connecting multiple assembly persons and senators from different districts.

---

7 Reliably identifying small businesses is quite difficult. Many individual proprietors or partners may list themselves as individuals. Determining the scope of a business and its ownership form is not always obvious. Generally, donors who explicitly list a business name, LLP, law offices of, or a business type that is almost always local in scope are coded as small businesses. It is possible that some of the entities identified are actually very large companies with wider geographical scope.

8 Identifying big businesses is less difficult than some of the other donor types, but still not perfectly reliable. Most in this category have widely recognized names (e.g. Pacific Gas and Electric Corp.). If anything, our coding may under-represent the big business category by placing some large scale and open ownership enterprises in the small business category. It is not likely that smaller closed-ownership forms have been mistakenly coded as big business.

9 We use the term 'investment-driven' rather than Clawson et al.'s 'pragmatic' because we believe it more accurately describes groups with instrumental motivations for contributing.

10 http://www.emilyslist.org/

11 http://www.thewishlist.org/

1.4 Hypotheses

Based on our above assumptions about donor motivations we can make the following hypotheses about the characteristics of candidates tied to the six donor types we've identified.

Individual citizens who donate are likely to make small donations that are driven by expressive motivations (Ansolabehere, de Figueiredo, & Snyder, 2003; Francia et al., 2003). Donations by individuals are likely to be strongly colored by ideological factors and therefore we hypothesize that individual donors in the candidate-donor network will make contributions that connect candidates from the same party. Similarly, we expect bias against individual donors connecting direct opponents. Because we categorize individual donors as relatively ideology-driven, we expect that a portion of individual donors will be biased towards connecting multiple minority and female candidates, whose shared interests they hope to promote. Individual donors should mostly tie together candidates that share similar geographic districts due to the localized interests of individuals. Therefore we might expect strong constituent overlap between candidates tied to individual donors. We also would only expect individual donors to be supportive of their own district's assembly person and senator. Therefore, we would not expect to find strong ties between multiple senators or assembly persons and individual donors.

Other than scale of donation, the pattern of donation of candidate committees should resemble those of individual donors with a strong ideological component. Therefore, we hypothesize that candidate committees will be strongly tied to multiple candidates of the same party, and disproportionately not connected to direct opponents. Candidate committees often contribute to multiple candidates in the same election cycle. Because they likely contribute to candidates with similar agendas, we might expect some candidate committees to center some of their support for candidates on demographic characteristics like ethnicity or gender. Though relatively local, candidate committees should be more likely to go beyond a specific region to build networks of candidates than will individuals but should have a small local bias (Gimpel et al. 2006, 2008).

Social organizations generally seek to provide representation of the membership class on a wide range of interests and seek institutional access. They probably display both ideology-driven and investment-driven campaign spending. We would anticipate that their behavior would favor established politicians, and yet that they would have a rather strong partisan bias in their investments. Similar to candidate committees, we expect social organizations to center some of their support for candidates on demographic characteristics like ethnicity or gender. We do not expect social organizations to be locally motivated in their campaign contributions.

Due to their similarity to social organizations, we might expect similar patterns of behavior regarding donation motivation for PACs. They will likely invest in established politicians, yet will have a strong partisan bias in their investments. Recent research has demonstrated that only a small percentage of PACs are "nonideological" in their campaign contributions (Bonica, 2013). It has been established that certain PACs such as Emily's List and the Wish List support multiple female candidates while others like Businesses Supporting Minority Candidates contribute to multiple minority candidates. Therefore, we expect PACs to be strongly connected to multiple female and minority candidates in the candidate-donor network. We do not expect PACs to be locally motivated in their campaign contributions.

Finally, based on findings from Clawson et al. (1986), we would expect that businesses would display both ideological and investment tendencies, with smaller businesses more likely to contribute ideologically by connecting candidates of similar parties while more interest-driven large businesses connect candidates of different parties. We would expect both large and small businesses to invest in incumbents and hedge by connecting political opponents. We do not expect either large or small businesses to be motivated by candidate demographics in their campaign contributions. We also hypothesize that small businesses will reflect local interests in their campaign donating, focusing on legislators that will directly influence business in their district.

2. Data and Methods

In order to test our hypotheses, we collapsed the two-mode candidate-donor network into a one-mode candidate-candidate network where the candidate nodes were said to be tied to one another via shared campaign donors. We built separate candidate-candidate networks for each type of donor. Below, we examine the candidate-to-candidate networks by donor type. We collected candidate attribute data in order to see how candidate attributes influence donation patterns by donor type in an effort to understand donor motivations. California provides the main empirical example for work on the importance of extra-legislative actors (Masket 2007, 2009; Koger et al 2009) and so we are addressing a key case in that literature. Also, we generally know less about state legislatures than the US Congress and so it seems reasonable and defensible to seek to extend our understanding into less explored

political arenas.

The choice of primary elections as a focus for analysis is relevant because pronounced district safety in many, if not most, US state elections means that the general election is often a foregone conclusion. This is especially true in California where the majority of general elections are considered "safe" (McCarthy, 2013). General elections tend to offer donors a take it or leave it choice of supporting an incumbent. Primary elections allow donors much more scope to "vote with their pocketbook". Focusing upon primary elections provides insight into donation patterns at a critical stage of the electoral process and is identified as a key stage in the literature on extra-legislative actors.

California's public records allowed us to gather appropriate data. We were able to construct a data matrix that links over 45,000 donors who made over 63,000 donations to 139 candidates who reported donations in the primary elections of 2004. Data on donors and donations were available from the web site of the Secretary of State of California after the end of the mandated reporting period for the 2004 primary election cycle. The raw data are posted exactly as reported by individual donors, and required substantial editing. As a first step, the lists of individual donations to individual candidates were combined into a master file. Next, extensive editing of the names of donors was conducted to assign a single standard name to all of the donations made by each contributor. Careful checking and cross-checking by multiple coders helped to ensure that the donations by multi-campaign donors were correctly identified.

Most of these donors are formal organizations, corporations, or small businesses, and most variations in the reported names of these entities could be reliably located. Our coding of donations by individual persons is believed to be rather less reliable. Individuals are more likely than formal organizations to use different names for reporting different donations (e.g. J. Smith versus J. M. Smith). When we could not be confident that a name variation was the same person, we assumed that it was separate persons. It is also more likely that there are multiple different individual donors with the same reported names. In many of these cases, other information (occupation, employer, address) could be used to resolve ambiguities. For individual donors, however, the coding is less than perfect. Fortunately, there is a strong tendency for individual donors to contribute to only one, or small numbers, of campaigns, and hence not affect our results on the patterns of donor network overlap very much. Failure to identify multiple donations by the same donor correctly may contribute to a downward bias in the intercept in the models in Table 3, mostly for individual donors. Errors of this type are probably not correlated with the attributes of candidates (e.g. the candidate's gender, ethnicity, location, district size, etc.). So, it is unlikely that there is related bias in model coefficients. After names of donors were standardized, all donations by the same donor to the same candidate were combined.

A matrix of donors who contributed to more than one candidate (4,746) by candidates (139) was constructed, with the dollar amounts of total contributions as cell entries, forming a two-mode or "affiliation" matrix (hanneman & Riddle, 2005). This matrix was then made binary (i.e. a donor did, or did not contribute to a given campaign). We decided to treat donations to candidates as present or absent, rather than retaining the amount of donations, since the focus of our analysis is on the numbers of donors in common between candidate pairs. The two-mode matrix was then used to induce a one-mode matrix of candidates by candidates (i.e. 139 by 139) showing the number of donors in common between each pair of candidates. This matrix serves as the dependent variable in the regression analyses reported in Table 3. Independent variables were similarly prepared as matrices describing the joint attributes of each pair of candidates. For example, the data for the independent variable *Republican* consists of a 139 by 139 matrix with each element coded "1" if both candidates are Republicans and "0" otherwise. Binary matrices of this type are prepared for all of the independent variables except the variable *Common Constituency*, which is an integer valued array and will be described in detail below.

To examine the effects of various factors on the amount of donor overlap in the networks of pairs of candidates, we regressed the matrix of numbers of donors in common for pairs of candidates on the matrices containing the variables describing attributes of the pair (e.g. were both candidates Republicans? How many constituents did the two candidates have in common?). The data are configured in a "round robin" design of dyads of candidates with multiple (but balanced) observations of each candidate (Kenny, et al. 2006). That is, the 19,182 pairs observed are composed from 139 individual candidates.

There are a variety of possible approaches to obtaining coefficient estimates for predictive models with dyadic data. We utilized the quadratic assignment procedure (QAP) algorithm in UCINET 6 (Borgatti, Everett, & Freeman, 1992). Significance tests for effects are calculated using the Y-permuation method. That is, many (in our case 2,000) runs of each model are made, randomly assigning scores on the dependent variable (numbers of donations to each candidate pair) to the vector of scores of each case on the independent variables.

The standard deviations of the distributions of parameter estimates from these random trials are used as estimates of the standard errors of coefficients. The hypothesis tests should be interpreted as tests of the null hypothesis that parameters are the result of random processes, given the observed distribution of data. Y-permutation significance tests do not test generalizability of the results to some larger population.

In the primary elections of 2004 in California we were able to identify 139 candidates who received one or more campaign contributions reported to the Secretary of State (several other candidates were on ballots, but did not report any contributions, and are not included in our analyses). Of these candidates, 117 were contenders in races for 64 seats in the Assembly (lower house), and 22 were running for 16 seats in the California state Senate. A substantial number of campaigns were not competitive. Of the 64 races for Assembly seats, 31 had only a single candidate (28 Democrats, three Republicans). Thirteen more of these contests had multiple candidates, but from a single party (seven Democratic, six Republican). Fourteen of the Assembly district primaries had a single candidate from each party; six had multiple candidates from one party but a single candidate from the other. No district had competition between multiple candidates of both parties. The level of competition was even lower in the Senate. Of the 16 races, 10 had single candidates (one Democrat, nine Republicans). Three districts offered competition within a single party (two Republican, one Democrat), and two districts had a single candidate from each party. In the Senate, as in the Assembly, there were no contests that featured multiple candidates from both parties.

We generated 14 variables to examine donor motivations. The variables *Democrat* and *Republican* indicate whether or not both candidates in a given pair are from the same major party. Ninety of the 139 candidates were Democrats, and 49 were Republicans. These variables are used to measure whether or not certain donor types tend toward ideology-driven contributions by connecting only candidates of a single party, or investment-driven contributions connecting candidates across parties. We include the variable *Opponent* to indicate whether each pair of candidates were running for the same office (whether their opponents were of the same party or not). Ideology-driven donors should be biased against contributing to opponents, while investment-driven donors should be biased toward this behavior.

The variable *Experienced* indicates whether or not a given pair of candidates consists of incumbents or candidates that have both previously held a state-level position. In contrast, the variable *Inexperienced* indicates whether both candidates are not incumbents or did not previously hold a state-level position. Of the 139 candidates, 56 were incumbents or prior state-level office holders (California has term limits in each house, and it is not uncommon for office holders to move from one chamber of the legislature to the other). Eighty-three of the candidates had not previously held state office. We assume that investment-driven donors are more likely to connect experienced candidates while ideologues should connect inexperienced candidates or challengers.

We argued that certain donor types are ideologically motivated to connect candidates of specific demographic characteristics such as gender and ethnicity. Among the 139 candidates analyzed here, 50 are female and 89 male. Four of the candidates are Black, eight Asian, 35 Hispanic, and 92 White. The following demographic variables, *Female*, *Male*, *Black*, *Asian*, *Hispanic*, and *White*, indicate whether or not gender or ethnicity is the same for both candidates in a given pair.

We employed three variables to measure local interest by donors: *Common Constituency*, *Senate*, and *Assembly*. The variable *Common Constituency* is the number of constituents that two candidates had in common which was available from California Senate re-districting web sources (California State Senate, 2007). It counts the number of registered voters in the overlapping districts of any two candidates. For two candidates competing for the same office, this number is the total registered voters in the district. In California, Assembly and Senate districts display highly variable degrees of overlapping constituencies. The variables *Senate* and *Assembly* indicate whether or not both candidates in a given pair are running for the same type of office. Of the total of 139 candidates, 117 are running for the Assembly and 22 for the Senate. These variables indicate local motivation in campaign donation patterns. Those donors that are more likely to generate strong ties between candidates sharing common constituencies and less likely to support multiple senators or assembly persons are considered more local in their scope of donating.

3. Results

3.1 Candidate-Donor Networks

First, we should examine the extent to which the relationships between candidates and donors are a network rather than a set of individual relationships. The second column of Table 1 displays the number of donations given by donors broken down by the number of campaigns the donor gives to. For instance, there are 41,043 donations given by donors that contributed to one

and only one campaign. There are 5,908 donations given by donors that contribute to two and only two campaigns, and so on. As seen in Table 1 the majority of donations are "one-offs," or donations that came from donors who have given to a single campaign only. This supports the view that donations are independent acts. But a large share of donations – and the majority of money donated – goes to multiple campaigns. Also, it is clear that there exists a significant number of "super donors" who give large amounts of money to multiple campaigns.

These figures show that for a significant proportion of all campaign donations and a majority of campaign finance, campaign donations are not independent acts between a single donor and a single candidate. The image of individual donors supporting a single candidate, or perhaps one candidate for each house of the legislature does characterize the majority of donors in primary elections. There are, however, a surprisingly large number of donations (12,553) that came from donors who supported more than five candidates in the election cycle. These campaign donors (about 10% of the total of all donors) strongly contribute to the web of overlapping constituencies; and they contribute almost half of all of the money.

Because of laws limiting the sizes of donor's donations to each candidate, the median sizes of contributions of donors to individual candidates do not display a great deal of variance (see final column of Table 1). Donors who contribute to multiple campaigns (and hence form overlaps between donor networks) do make markedly higher contributions to each campaign than "one-off" donors. Because "super-donors" contribute in very large numbers of campaigns, there is considerable variation in the total expenditures of individual donors in the election cycle as a whole.

Table 2 provides descriptive information by type of donor for those donors who contributed to more than a single campaign (4,746 of 45,802 donors). The largest numbers of donors who link the campaign contribution networks of candidates are individual persons and small businesses. They are most likely to make small contributions and participate in only two campaigns. Because of their large numbers, they do provide a significant amount of the money linkage among campaign networks. Institutional and PAC donors link more campaigns, make larger contributions, and contribute the largest share of all of the money linkages among candidates. Big business donors contribute to the largest number of campaigns, and individually spend more money than other types. Since there are relatively few such donors, however, their contribution to the overall financial linkages among candidates is modest. Contributions among politicians occurred in relatively small networks, and were relatively few in number. The amounts of money flowing in these channels, however, were significant. Though there are many single campaign donations, we can conclude from these tables that conceiving of donations as independent from one other is a simplification that diminishes descriptive accuracy.

3.2 Donor Motivations

Beyond establishing that SNA provides a new description of donation activity our major expectations concerned the patterns of linkages that speak to donor motivations. The patterns in the candidate-candidate networks allow us to identify the ways in which different types of extra-legislative actors donate money and so allows us to speak to broader arguments about the role of extra-legislative actors. Table 3 reports regression models for the number of donors that each pair of candidates (i.e. 9,591 unique pairs formed between 139 candidates) have in common.

Table 1: Donations to Assembly and Senate candidates by the number of campaigns to which contributions were made.

Number of Campaigns	Donations	Donations (%)	Total amount donated ($1,000)	Median donation per candidate
1	41,043	64.8	35,118	$250
2	5,908	9.3	6,091	$400
3	1,974	3.1	3,484	$500
4	1,052	1.7	1,564	$1,000
5	765	1.2	1,297	$1,000
6-10	2,473	3.9	7,310	$1,000
11-20	3,017	4.8	10,990	$1,250
21-40	3,656	5.8	8,260	$1,500
41-60	2,417	3.8	8,194	$2,000
61-82	990	1.6	3,349	$3,000
Total	63,295	100.0	85,658	$500

Table 2: Multiple-campaign donors to Assembly and Senate candidates by donor type.

Donor type	Donors	Median donation	Median campaigns	Median donation per candidate	Total amount donates ($1000)
Individuals	2,358	$789	2	$250	$5,662
Candidate committees	220	$9,600	4	$3,200	$3,601
Social organizations	186	$10,225	6	$2,000	$8,093
PACs	520	$6,950	4	$1,425	$19,306
Small business	1,346	$2,600	2	$1,000	$9,492
Big business	116	$18,550	12	$1,500	$4,386

Separate models are reported for each donor type.

The intercept is the predicted number of co-donors of a hypothetical pair of candidates who are not competing against one another and who have no constituents in common. One of the hypothetical pair has prior office experience and the other does not; one is running for a senate seat, the other for the assembly; one is a Democrat, and the other a Republican; one is male, the other female; the two candidates are of different ethnicities.

From Table 3, we can see that the predicted number of shared donors between any two randomly chosen campaigns, net of other factors, is very small among individual donors and candidate committees. That is, these two types of contributors are relatively less likely to link the donation networks of multiple candidates. Social organizations, PACs, and business donors are much more likely to play linking roles.

When interpreting the meaning of the other coefficients, it is useful to keep the differences in the overall mean numbers of donor overlap in mind (means are shown near the bottom of Table 3, and display patterns very similar to the intercept values). For example, an effect of the same absolute size for the individual donors network and PACs network is much more substantively "important" for the individual donors network, because it represents a much larger effect relative to the typical number of shared donors in that network.

The findings related to donor motivations for individuals and candidate committees were very similar so we start by discussing them together. We argued above that donor motivations might fall on a spectrum between ideology-driven and investment-driven agendas. Donors that contribute for ideological purposes are likely to support one political party or the other rather than to spread their contributions across parties. Consistent with our hypotheses, individuals and candidate committees are the strongest supporters of partisanship by favoring to support candidates of the same party (see Table 3). For example, in individual donor networks, two Republican candidates have almost four times as many ties ((0.199 + 0.573) / 0.199 = 3.88) and Democrats almost four and half times as many ties ((0.199 + 0.695) / 0.199 = 4.49) as a mixed party pair.

Another indicator of donor motivation which demonstrates the investment motive of donors is whether or not they support opposing candidates. We argued that donors supporting candidates that are running against one another in the same race are likely hedging their investments in the race and ensuring that they have contributed to the winning campaign. As expected, we find that a bias against contributing to both of two competing candidates exists for individuals and candidate committees.

We also examined whether or not donors invest in pairs of experienced candidates (or avoid donating to multiple inexperienced candidates) as a means of buying access to incumbents or experienced politicians who are typically favored in state legislative elections (Carey, Niemi, & Powell, 2000). Here, we see the motives of individual donors and candidate committees diverge. Contrary to our hypothesis, we find that individuals strongly support connecting experienced candidates. However, this is less surprising when we consider that a popular explanation for the advantage of incumbency is that individuals tend to approve of the local activities of their own elected representatives, even while disapproving of the larger bodies of legislature (Cillizza, 2013; Cook, 1979). While the findings for candidate committees are not significant, it should be noted that they are the only donor type that tend to tie together inexperienced candidates at a higher rate than experienced candidates.

Due to similarity among certain minority groups in terms of the political issues that concern them, we should expect donor support for multiple women and ethnic minority candidates by individuals and

Table 3: QAP regression models predicting number of donors in common among candidate pairs by types of donors to Assembly and Senate candidates.

Variables	Individuals	Candidate Committees	Social Organizations	PACs	Small Businesses	Big Businesses
Motivation						
Republican	0.573**	0.904**	0.269	1.751**	1.225**	0.804*
Democrat	0.695**	0.968**	3.125**	2.078**	0.164	-0.744*
Opponents	-3.082**	-1.081**	-1.973**	-1.226	1.764**	0.532
Experienced	0.100*	0.039	8.373**	13.870**	8.661**	8.348**
Inexperienced	0.018	0.060	-2.125**	-3.455**	-1.779**	-1.711**
Localism						
Common Const.[a]	0.028**	0.002*	0.003*	0.001	0.007**	-0.002
Senate	0.005	0.296*	-1.084**	-0.476	0.259	-0.060
Assembly	-0.318**	-0.016	0.445	-0.294	-0.377	-0.397
Demographic						
Female	0.473**	0.940**	0.022	0.209	-0.379	-0.445
Male	0.025	-0.337**	0.043	0.062	0.493*	0.376
Asian	1.880**	-0.402	0.401	1.640	1.011	0.005
Black	0.046	-0.460	4.247**	-0.423	4.082**	-0.980
Hispanic	0.025	0.565**	1.358**	3.748**	2.926**	3.551**
White	-0.027	-0.173*	-0.714*	-1.742**	-1.192**	-1.247**
Intercept	0.199	0.134	2.163	5.510	3.119	3.086
Mean	0.057	0.612	4.178	6.901	3.691	3.092
S.D.	2.438	2.180	6.134	9.338	5.753	5.730
R^2	0.103	0.093	0.472	0.471	0.435	0.400
N=19,182 pairs						

[a.] units of 1,000 constituents

* $p < 0.10$, ** $p < 0.05$, one-tail

candidate committees who hope to advance similar political agendas. We find support for our hypotheses in the candidate-candidate networks. In Table 3, we note a significant tendency for individual and candidate committee donors to make contributions to candidate pairs who are both female. Candidate committees provide eight times the amount of ties per pair to female-female pairs of candidates than to mixed gender candidate pairs. They are also less likely to connect pairs of male candidates than to connect mixed gender pairs. This demonstrates that candidate committees are a strong form of political support for female candidates running for the state legislature. There is very strong donor overlap for Asian-American candidates among individual donors. This is consistent with findings by Cho (2002) that Asian American contributions are strongly tied to ethnicity. Individual donors provide over ten times the amount of ties per pair for pairs of Asian candidates than for mixed ethnicity candidates. Also, strong donor overlap from candidate committees exists for Hispanic candidates.

To test for effects of locally motivated donors, we examine the coefficients for the *Common Constituency*, *Senate*, and *Assembly* variables. If donor motivations are relatively local in scope, we should see bias towards ties between candidates that have a lot of constituents in common with each other (net of other important covariates like whether or not the two are opposing each other in the same race). The regression coefficients for *Common Constituency* display the number of increased shared donors for each additional increase of 1000 shared constituents. However, the magnitude of the coefficients are relatively small (compared to the mean ties for each donor type) for all of the donor types except individual donors where it is fairly large. This suggests that there is a strong geographical component to donation motives of individual donors, where as expected, they tend to be more local in scope tying candidates from similar geographic regions together. The effect of tying together candidates with overlapping constituency is significant for candidate committees as well, but as expected, their

interests are not nearly as local in scope as they are for individuals. We hypothesized that individual donors, with their strong local orientation, would not support multiple Senator or multiple Assembly candidates instead supporting one of each from their district. We find that there is a relatively strong bias against supporting pairs of Assembly candidates by individual donors, but no such bias for or against supporting a pair of Senate candidates. Surprisingly, we find that candidate committees are over three times as likely to connect Senators as they are to connect mixed legislator pairs.

The motives of social organizations are a little more complex. As we expected, they do display some partisanship indicating ideological motives. However, they are only significantly biased in favor of connecting Democrats with twice as many ties to Democratic candidate pairs than to Republican or mixed party pairs of candidates. This actually isn't surprising given the presence of unions among this donor type. These types of donors are resistant to connecting direct opponents, but are heavily invested in tying together experienced candidates with a bias against inexperienced candidates. Due to mobilization surrounding the disproportionate representation of women and ethnic minorities in politics, we argued that we expect patterns of support among social organizations for these underrepresented candidates. Though we do not find social organizations to be motivated by the gender of the candidates they support, we do find that social organizations are the strongest supporter of multiple Black candidates connecting Black candidates via campaign contributions nearly three times as often as pairs of candidates of differing races. There is also significant support in favor of connecting Hispanic candidates by social organizations as well. There is small significant support for connecting candidates with overlapping constituency by social organizations, along with significant bias against supporting multiple Senators indicating more local oriented interests by social organizations than we hypothesized.

The motives of PAC donors are mostly consistent with our hypotheses. They are ideologically driven and are significantly more likely to connect pairs of single party candidates of both major parties. However, as expected they tend to strongly invest in connecting experienced candidates while simultaneously avoid connecting inexperienced candidates. We expected PACs to strongly support underrepresented candidates and even cited examples of specific PACs whose mission it is to do just that. We were surprised to find weak non-significant support for connecting female candidates by PAC donors. However, we did find significant support for connecting Hispanic candidates. There were no significant effects for the local interest variables as expected.

It is worth directly comparing the motivations of small and large business donors. Both small and large businesses display partisan ideology in favor of Republican candidates, but the effect sizes are relatively small. Small businesses are only 39% more likely to connect pairs of Republican candidates than mixed party candidates, and large businesses are only 26% more likely to connect Republican pairs. Large businesses are also significantly less likely to connect Democratic candidates via campaign contributions. Both small and large businesses are more likely to connect direct opponents than two candidates that are not directly running against one another, however, the effect is only significant for small businesses. This supports the hypothesis that business donors are indeed investment-driven when it comes to hedging their campaign investments. Also, both types of business donors strongly support connecting experienced candidates and avoid connecting inexperienced candidates. In general it appears that both small and large business donors are relatively investment oriented, however, we also predicted that small businesses would be more driven to connect single party candidates than large businesses, which doesn't appear to be the case. We did not expect business donor behavior to be motivated by candidate demographics, but found that both types of business donor are biased toward connecting Hispanic candidates, and small businesses were also biased toward connecting Black candidates. We also find that both small and large businesses display relatively small tendencies to support male-male candidate pairs and discourage female-female pairs. As expected, small businesses also frequently connect candidates with overlapping constituents which reflects their local interests.

In general, we find that most donor types tend to prefer to participate in campaigns of multiple candidates of the same party. For some donor types there is a stronger Democratic partisan bias (among individual, social organizations, and PAC donors) with social organizations demonstrating the largest disparity in support favoring democratic pairs. Meanwhile, business donors tend to connect Republican candidates. From these results it appears that all donor types display some ideological motivations by supporting or not supporting shared party pairs as compared to mixed party pairs. However, individual and candidate committees are the strongest promoters of political partisanship via their campaign donations, while business donors appear to be the most investment-driven. These findings also indicate that all of the donor types except candidate committees tend to invest in incumbents by preferring to connect experienced candidates through campaign contributions

while avoiding spreading their money to multiple inexperienced candidates.

We also found, unexpectedly, that if both members of a pair of candidates are white, most donor types display a relatively small but significant preference against the pair (compared to a pair of candidates of different ethnicities). Therefore there doesn't appear to be a tendency to support multiple candidates due to their white ethnicity net of covariates. This is not surprising considering the overrepresentation of white politicians in the United States (Lublin, 1997; Malhotra & Raso, 2007).

4. Discussion

On the basis of our findings we can draw several conclusions. First, simply at a descriptive level, we have shown that there are indeed networks of candidates linked by extra-legislative actors: donations are not individualistic but are embedded in a wider context. This not only provides a new way of representing the relationships between donors and candidates it also offers a way of understanding some of those relationships.

We also found that there is complexity in donor motivation that depends on the type of donor. No single donor type is found to be exclusively ideologically or investment driven. We generally conclude that individuals and candidate committees tend to be more ideologically driven, business donors tend to be more investment driven, and social organizations and PACs fall somewhere in between.

Masket (2007) claims that ideological extra-legislative actors are often the drivers of polarization in legislatures. The segment of extra-legislative actors that controls resources candidates depend on to get into office (i.e. donors) act as "gatekeepers to public office" (p. 484). If candidates and elected officials want continued support from donors, they must satisfy donor interests. These actors are known to influence the behavior of elected officials regardless of public opinion (ibid). Therefore, according to Masket, if donors reflect partisan behavior in their contributions to candidates, their ideology is likely to be reflected by the candidates. Further, since fundraising is core to the electoral process, with most state level legislators spending at least a quarter of their time in office fundraising (Hernnson, 2008), candidates and elected officials that are tied financially by donors are also more likely to connect with one another at fundraising events. When donors are biased in favor of linking single party candidates, it may well reduce the likelihood of bipartisanship among elected officials. Ultimately, patterns in our data help to support Masket's argument that extra-legislative actors have a relationship to campaigns that promotes an ideological division. To varying degrees, all of the donor types tended to prefer donating to multiple candidates from the same party.

By mostly contributing to candidates from parties with similar ideology, candidate-donor networks should emerge wherein candidates of different parties are mostly isolated from one another. To the extent that extra-legislative actors both align with and help support party division, we would expect to see that party primarily determines networks. For example, in the U.S., we should see network patterns strongly shaped by party label and, in effect, see two distinct candidate-donor networks (one Democrat, one Republican) with very few connections between the two.

When we examine the core of the overall candidate-donor network by collapsing to a single mode candidate-candidate network (this network includes donors of all types acting as ties; see Figure 1), we find that while some of these relationships are distinctly partisan many are not. Figure 1 shows the network core, or most densely connected set of candidates. Here, we define "core" as the set of candidates that share at least 100 donors in common with at least one other candidate (this is represented with a tie or linkage in the graph). Graphical representations of polarized politics typically show a cluster of Republican candidates on one side of the figure and a cluster of Democratic candidates on the other side with very little or no connection between them. To some extent this is also the case here with the horizontal dimension being predictive of party (Republicans are found on the left side of Figure 1 while Democrats are on the right). However, we also see that a number of the Republicans and Democrats are connected in the sense that they have multiple donors in common; e.g. Garcia, a Republican, and several other Democrats including Horton and Machado. What this figure and our previous results suggest is that there are different kinds of donor motives at work. The strong tendency to tie candidates of the same party together by most donor types is moderated by the preference for connecting experienced candidates creating the network core found in Figure 1.

In the analysis in Table 3, we identified the "experience" of candidates as quite an important attribute in driving donations. Like partisan bias, we found that preference for connecting experienced candidates in candidate-donor networks is commonplace (every donor type except for candidate committees demonstrated significant bias in favor of connecting experienced candidates or against connecting inexperienced ones). If donors to multiple campaigns are investing by connecting experienced candidates, then we'd expect to find highly connected incumbent candidates in the dense core of candidate-donor networks while challengers would have

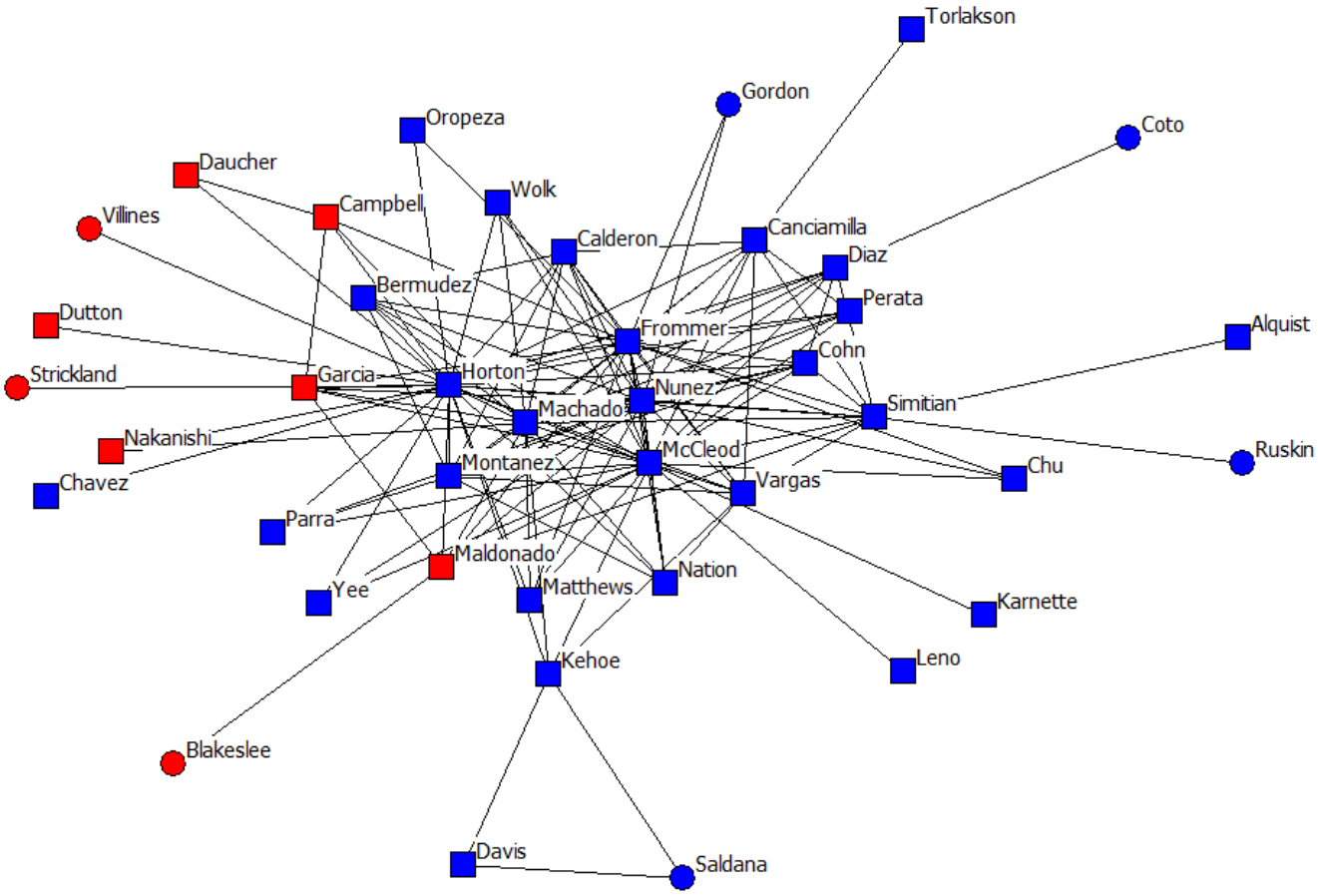


Figure 1: Network of highly connected candidates (each tie between candidates indicates at least 100 donors in common).
Note: The spatial arrangement is a spring embedding algorithm from UCINET's netdraw. It is a two-dimensional MDS, but with locations modified somewhat to minimize overlap of nodes, and minimize crossing of lines.

Key:
Democrat - blue, Republican - red
Experienced - square, Inexperienced - circle

fewer connections and be relegated to the fringe of the network. Ultimately, this is apparent in the core of the single mode candidate-candidate network. In Figure 1, experienced candidates are shown as squares. We note that the most highly interlinked candidates are likely to be experienced and that often these linkages cross partisan boundaries; i.e. incumbents share ties across the party divide and take money from at least 100 of the same extra-legislative actors.

Does centrality in the common-donor network matter? The ultimate test, of course, is winning elections. In our data, the association between centrality in the common-donor network and victory in the primary election is a strong one (r = 0.82). While one cannot directly attribute victory to the success of candidates in embedding themselves in dense donor networks due to the high correlation with incumbency, the magnitude of the relationship is remarkable. This suggests that donor-network centrality may be part of the process by which winners win, and losers lose. We believe this finding can contribute to existing theories of incumbency advantage (Cary et al., 2000).

By looking at the core of the overall donor network (Figure 1), it is apparent that sizeable networks of relationships work to encourage ties only among Democrats or only among Republicans. These patterns are consistent with the argument of Masket to the effect that extra-legislative actors can, through their behavior, support polarization. However, it should be noted that in the core of the donor network, the party divide is bridged by well-connected experienced candidates.

The interpretation of donation motives driven by candidate demographics is less clear-cut. All donor types were found to significantly connect some demographic characteristics, with motivations to connect Hispanic candidates and bias against connecting white candidates being almost universal. This is likely a reflection of the demographics of California, where efforts are being

made to create a legislature that is more representative of the growing Hispanic population. Women and Asian-American candidates seem to rely on individual donations more fully than do Anglo, Latino, and Black men: this would imply that their candidacies face more difficulties in gaining access to deeper pockets than others and that the instincts of groups such as Emily's List are well-founded.

Finally, we found that many donor types are driven by local interests including those we expected to be: individuals, candidate committees, and small businesses. However, we found social organizations were also locally oriented though the effects were not nearly as large as they were for individual donors. There doesn't appear to be any general or consistent bias towards or against connecting a particular level of legislature as can be seen by the senate and assembly variables in Table 3. There also appears to be a complex relationship between having political experience, being embedded in candidate-donor networks, and winning campaigns. It is impossible to determine causality given our research design, but it is clear that incumbents and prior holders of state level offices are better connected to each other via donors and are ultimately more likely to win elections. This is likely due to multiple feedback processes between these various phenomena and should be investigated in further research.

By taking a network approach to understanding campaign contributions, we are able to identify behavioral patterns associated with motivations by donor type using the actual pattern of donation that ties candidates together. This approach treats donor behavior as linked across candidates rather than separated by candidate. The SNA approach focuses attention on the processes that create political structures – shared interests, similarities of stakeholders, and potential legislative cooperation among candidates. Seeing both candidates and donors as "embedded" in, and creating structures linking voters and candidates creates new insights into the role of money in politics.

References

Ansolabehere, S., de Figueiredo, J. M., & Snyder, J. (2003). Why is there so little money in U.S. politics? *Journal of Economic Perspectives*, 17(1), 105-130.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2, 113-120.

Bonica, A. (2013). Ideology and interests in the political marketplace. *American Journal of Political Science*, 57(2), 294-311.

Borgatti, S., Everett, M., & Freeman, L. (1992). *UCINET IV Version 1.0 User's Guide.* Columbia, SC: Analytic Technologies.

Bowler S. & Hanneman, R. (2006). Just how pluralist is direct democracy? The structure of interest group participation in ballot proposition elections. *Political Research Quarterly*, 59(4), 557-568.

Bratton, K. & Haynie, K. (1999). Agenda setting and legislative success in state legislatures: The effects of gender and race. *The Journal of Politics*, 61(3), 658-679.

Brown, C., Powell, L., Wilcox, C. (1995). *Serious money: Fundraising and contributing in presidential nomination campaigns*. Cambridge University Press.

Burris, V. (2001). The two faces of capital: Corporations and individual capitalists as political actors. *American Sociological Review*, 66(3), 361-381.

California State Senate. (2007). Reapportionment Retrieved from http://www.sen.ca.gov/ftp/SEN/COMMITTEE/STANDING/EL/_home/Reapportionment/reapportionment.HTP.

Cary, J. M., Niemi, R. G., & Powell, L. W. (2000). Incumbency and the probability of reelection in state legislative elections. *Journal of Politics*, 62(3), 671-700.

Cho, W. T. (2002). Tapping motives and dynamics behind campaign contributions. *American Politics Research*, 30(4), 347-383.

Cho, W. T., & Gimpel, J. G. (2007). Prospecting for (campaign) gold. *American Journal of Political Science*, 51(2), 255-268.

Cillizza, C. (2013, May 9). People hate Congress. But most incumbents get reelected. What gives? The Washington Post. Retrieved from http://www.washingtonpost.com/blogs/the-fix/wp/2013/05/09/people-hate-congress-but-most-incumbents-get-re-elected-what-gives/

Clawson, D., Neustadtl, A., & Bearden, J. (1986). The logic of business unity: Corporate contributions to the 1980 congressional elections. *American Sociological Review*, 51(6), 797-811.

Coffey, D. (2007). State party activists and state party polarization. In J. Green and D. Coffey (Eds.), *The state of the parties: The changing role of contemporary American politics*. Lanham, MD: Rowman & Littlefield.

Cohen, M., Karol, D., Noel, H., & Zaller, J. (2008). *The party decides*. Chicago: University of Chicago Press.

Conti, J. P. (2002). The forgotten few: Campaign finance

reform and its impact on minority and female candidates. *Third World Law Journal*, 22(1), 99-162.

Cook, T. E. (1979). Legislature vs. legislator: A note on the paradox of congressional support. *Legislative Studies Quarterly*, 4(1), 43-52.

Fowler, J. H. (2006). Legislative cosponsorship networks in the US House and Senate. Social Networks, 28(4), 454-465.

Francia P., Green, J., Herrnson, P., Powell, L., & Wilcox, C. (2003). The financiers of congressional elections. New York: Columbia University Press.

Gimpel , J., Lee, F. E., & Kaminski, J. (2006). The political geography of campaign contributions in American politics. *The Journal of Politics*, 68, 626-639.

Gimpel, J., Lee, F. E., & Pearson-Merkowitz, S. (2008). The check is in the mail: Interdistrict funding flows in congressional elections. *American Journal of Political Science*, 52(2), 373-394.

Gordon, S. B. (2001). All votes are not created equal: Campaign contributions and critical votes. *The Journal of Politics*, 63, 249-269.

Grant, J. T., & Rudolph, T. J. (2004). Expression vs. equality: The politics of campaign finance reform. Columbus, OH: Ohio State University Press.

Hall, R. L., & Wayman, F. W. (1990). Buying time: Moneyed interests and the mobilization of bias in congressional committees. *The American Political Science Review*, 84(3), 797-820.

Hanneman, R., & Riddle, M. (2005). Introduction to network analysis. Retrieved from the University of California website: http://faculty.ucr.edu/~hanneman/nettext/.

Herrnson, P. S. (2008). Congressional elections: Campaigning at home and in Washington. Washington, D.C.: CQ Press.

Jacobson, G. (1980). *Money in congressional elections*. New Haven: Yale University Press

Kenny, David A., Deborah A. Kashy, and William L. Cook. 2006. Dyadic Data Analysis. New York. The Guilford Press.

Knoke, D., & Kuklinksi, J. (1982). *Network Analysis*. Newbury Park, CA: Sage.

Koger, G., Masket, S. & Noel, H. (2009). Partisan webs: Information exchange and party networks. *British Journal of Political Science*. 39, 633-653.

Lawless, J. L. (2004). Politics of presence? Congresswomen and symbolic representation. *Political Research Quarterly*, 57(1), 81-99.

Levendusky, M. (2009). [Review of the book No middle ground: How informal party organizations control nominations and polarize legislatures, by S. Masket]. *Public Opinion Quarterly*, 73(4), 833–838.

Lublin, D. (1997). *The Paradox of Representation: Racial Gerrymandering and Minority Interests in Congress.* Princeton NJ: Princeton University Press.

Malbin, M. (Ed.). (2006). *The Election After Reform: Money, Politics and the Bipartisan Campaign Reform Act*. Lanham MD: Rowman & Littlefield.

Malhotra, N., & Raso, C. (2007). Racial representation and U.S. Senate apportionment. *Social Science Quarterly*, 88(4), 1038-1048.

Masket, S. (2007). It takes an outsider: Extra-legislative organization and partisanship in the California Assembly, 1849-2006. T*he American Journal of Political Science*, 51, 482-497.

Masket, S. (2009). N*o middle ground: How informal party organizations control nominations and polarize legislatures*. Ann Arbor: University of Michigan Press.

McCarthy, Devin. (2013, November 26). Did the California Citizens Redistricting Commission Really Create More Competitive Districts? *FairVote Research Reports*. Retrieved from http://www.fairvote.org/research-and-analysis/blog/did-the-california-citizens-redistricting-commission-really-create-more-competitive-districts/

McPherson, M., Smith-Lovin, L., Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 415-444.

Otte, E., & Rousseau, R. (2002) Social network analysis: A powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6), 441-453.

Panagopoulos, C., & Bergan, D. (2006) Contributions and contributors in the 2004 presidential election cycle. *Presidential Studies Quarterly*, 36(2), 155-172.

Peoples, C. D. (2008). Inter-legislator relations and policy making: A sociological study of roll-call voting in a state legislature. *Sociological Forum*, 23(3), 455-480.

Peoples, C. D. (2010). Contributor influence in congress: Social ties and PAC effects on U.S. House policymaking. *The Sociological Quarterly*, 51(4), 649-677.

Peoples, C. D., & Gortari, M. (2008). The impact of campaign contributions on policymaking in the

U.S. and Canada: Theoretical and public policy implications. *Research in Political Sociology*, 17, 43-64.

Siegel, D.A. (2009). Social networks and collective action. *American Journal of Political Science*, 53(1), 122-138.

Snyder, J. M. Jr. (1990). Campaign contributions as investments: The U.S. House of Representatives, 1980-1986. *The Journal of Political Economy*, 98(6), 1195-1227.

Stratmann, T. (2002). Can special interests buy congressional votes? Evidence from financial services legislation. *Journal of Law and Economics,* 45, 345-374.