Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Clearing the FOG: Fuzzy, overlapping groups for social networks

George B. Davis*, Kathleen M. Carley

CASOS, ISRI, SCS, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15223, USA

ARTICLE INFO

Keywords:
Groups
Stochastic models
Link analysis

ABSTRACT

Humans are well known to belong to many associative groups simultaneously, with various levels of affiliation. However, most group detection algorithms for social networks impose a strict partitioning on nodes, forcing entities to belong to a single group. Link analysis research has produced several methods which detect multiple memberships but force equal membership. This paper extends these approaches by introducing the FOG framework, a stochastic model and group detection algorithm for fuzzy, overlapping groups. We apply our algorithm to both link data and network data, where we use a random walk approach to generate rich links from networks. The results demonstrate that not only can fuzzy groups be located, but also that the strength of membership in a group and the fraction of individuals with exclusive membership are highly informative of emerging group dynamics.

1. Introduction

Since the earliest days of social network research, accurate detection of cohesive group entities has been an attractive and elusive goal. Group structure can be used for high-level descriptions of complex networks, to support or contest theories about underlying processes influencing social interactions, and to detect strengths or vulnerabilities of social structures and individual positions in a variety of contexts. These goals have important applications in a wide range of fields, including anthropology, sociology, organization science, economics, management, and security and intelligence programs.

Typically, group detection has consisted of dividing nodes into discrete partitions indicating mutual association. However, common sense and empirical analysis (Freeman, 1992) support the view that humans are capable of simultaneously filling many roles in many contexts, such that a strict partitioning may prevent detection of the true group entities in a graph. To better understand modular structure in networks, we must develop models which allow for multiple memberships and varied levels of membership.

In this paper we build off several link analytic group detection methods, due to Kubica et al. (2003b) and Battacharya and Getoor (2004), which allow for relaxed partitioning by permitting individuals under certain conditions to participate in multiple groups. We refine the representation of group structure by permitting varying strengths of association from members to group entities, and present an algorithm that generates such groupings from link data using a stochastic model of link emission from group entities and a maximum-likelihood clustering method. To analyze the utility of the fuzzy overlapping group model, we make comparison to groupings by anthropological observations and prior algorithms. Our results suggest that this approach is capable of identifying groups that are confirmed by existing quantitative methods as well as expert ethnographic analysis, while providing additional information about overlap between groups and individuals who play multiple roles. This additional information facilitates understanding emergent behavior in the groups.

The remainder of the paper is organized as follows. In Section 2, we provide a brief background on existing group detection methods. We also discuss the generation of link data from networks, so that we can apply our link analytic method to network datasets. In Section 3 we describe our approach (termed FOG for "Fuzzy Overlapping Grouper") in two subsections: one proposing a stochastic model of the way groups generate link data, and another introducing a corresponding maximum-likelihood method for inferring groupings based on evidence. In Section 4, we present performance results on the FOG algorithm and use FOG to analyze two well-studied real-world datasets: Sampson's monastery survey data (1968) and Davis, Gardner and Gardner's southern women (1941), comparing our results to previous groupings on the same. In the conclusion, we discuss FOG's potential contribution to group analysis based on our results, and identify additional work necessary.

2. Background and related work

2.1. Defining "Group"

Theoretically, we consider a group a set of entities which experience the same membership relation with respect to the same

[*] Corresponding author. Tel.: +1 412 580 3790.
E-mail address: gbd@cs.cmu.edu (G.B. Davis).

0378-8733/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.socnet.2008.03.001

external entity, real or abstract. In the social sphere, this can take many forms. For example: a formal organization like a board of directors, an implicit organization like a circle of friends, a demographic quality such as hair color, or even the set of individuals uniquely affected by an external force, such as the victims of a flu epidemic.

This paper is concerned specifically with *cohesive groups*, divisions which exhibit more associations within groups than between them. This operational definition may at first seem to restrict the varieties of group we can detect, yet we can imagine interaction data in which each of the above categories of group would leave such a trace. For example, members of a board of directors might occur together on the recipient list for formal memos and meeting announcements. Individuals afflicted by the same communicative disease might tend to be clustered in space and time in hospital records. To some extent, measuring this definition of cohesion depends on being able to clearly measure both the presence and absence of links between entities – a property inherent in social network data, but less obvious in link data, which we define and discuss in the next section. In link data, a stochastic model must fill the roll of defining what comprises a concentration of links. In the next section we describe several widely used algorithms to detect cohesive groups in both types of data.

Several non-cohesive group types are also popular in sociological research, particularly structural similarity. Entities are grouped together as structurally similar if their interaction patterns are similar; that is, if they interact with the same other entities or classes of entities. The group they experience membership with in this case is called a structural role. An intuitive example of a structural role is the middle manager in a hierarchical organization, who interacts with both upper management and employees, but not necessarily other middle managers. Like cohesive groups, structural roles can be implicit and unnamed or encoded in formal relationships. Like cohesion, the concepts of structural similarity and roles have been operationalized in many ways, the most common discovery techniques including block modeling (Lorrain and White, 1971) and CONCOR (Breiger et al., 1975).

Although FOG is designed to illuminate cohesive groups, we believe some researchers interested primarily in structural roles may be able to apply it to their study. Some correlations exist between structurally similar and cohesive groups. Membership in a strongly cohesive group can directly constitute a structural role, as interaction with other members dominates individuals' interaction patterns. Roles whose members do not interact can in some cases nonetheless be detected as a cohesive group following a transformation in data. For example, consider a dataset collected from vendors in which each lists the clients they sell to. If one inverts the association data to link vendors who share a common client, cohesive groups in this new dataset will collect vendors who filled similar roles in the original data. Finally, as we will discuss in our analysis of Sampson's data, detecting overlapping cohesive groups permits detection of a type of structural role, the interstitial actor.

2.2. *Finding cohesive groups*

One major theme in identifying cohesive network clusters has been an evolving series of graph theoretic group definitions, generally subgraphs satisfying internal connectivity requirement (e.g. cliques, *k*-clans, *k*-cores, *k*-plexes). These structures may overlap, leading to multiple community memberships. Palla et al. (2005) give an iterative definition, related to *k*-cliques, designed specifically to examine overlapping communities. Because group membership under their technique is binary, all individuals in a community overlap have equivalent positions. We hope that FOG can shed light on distinctions in these interstitial roles by adding weights of membership.

An alternative graph theoretic approach due to Moody and White (2003) emphasizes paths, defining cohesive communities as those supporting redundant communication threads. The same path-based rationale supports the heuristic method of Girvan and Newman (2002), which calculates communities by iterative removal of high-betweenness edges. Both techniques assign binary memberships, and those using the Newman's method lack capacity for overlapping or nested groups. Moody and White's communities do not overlap at any given level (FOG's do), but their hierarchy provides nesting relationships and is in some ways more informative as it supports the pairwise query: at what level are two individuals grouped? Newman's algorithm has seen several extensions and applications (Clauset et al., 2004; Newman, 2004a,b; Newman and Girvan, 2004) demonstrating its effectiveness on extremely large datasets which cannot be analyzed by other techniques (including FOG).

Another line of grouping research, block modeling, revolves around partitioning matrices such that subgroups have consistent relations. Block models are a natural setting for detection of structural equivalence (Lorrain and White, 1971), but have been extended to a variety of other settings including detection of cohesive groups (Doreian et al., 2005). Popular algorithms for block modeling include CONCOR (Breiger et al., 1975) and FACTIONS search (Borgatti et al., 2005). Recently, Doreian et al. (2005) have generalized block modeling for the analysis of 2-mode data, such as the relation of individuals and parties attended in Davis' study described below. Such relations are represented as rectangular matrices, and are block modeled by providing a separate partitioning for rows and columns. This closely relates to H-FOG's method, which discretely partitions one mode (the events). Rather than partitioning the other mode (individuals), however, FOG optimizes and presents fuzzy groups induced by the first partitioning.

In this paper, we often discuss 2-mode as described above, but refer to it as *link data*. This is to reference link analytic literature FOG is related to, to distinguish it from network data, and to emphasize FOG's perspective that we observe a sample of an infinite stream of links rather than the entirety of a finite matrix. We refer to our observed links as *evidence*, represented as an unordered set of *links*. Each link is an unordered set of entities in which each entity is assumed to have the same relation to an observation, such as "signed meeting roster", or "was observed in photograph". Our set of links may carry redundant associations (*i.e.* two events with same attendees) or simultaneous associations of more than two entities (one event with five attendees).

Data mining communities have produced several methods for extracting group entities from this type of data, including the GDA model/*k*-Groups algorithm (Kubica et al., 2003b) and Battacharya and Getoor's (2004) iterative deduplication method. These algorithms partition link data to infer groups which maximize the likelihood of observing the given data, according to a stochastic model. The fact that groups are built by partitioning links (not individuals) produces the advantage that, as with Palla et al., individuals may belong to more than one group. The method we introduce in this paper extends on these methods by allowing varying levels of association from entities to groups. This relaxation is intended to allow our group models to more tightly fit the data and to represent a wider variety of associative structures.

Another existing technique with some similarity to the FOG framework is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a recently introduced stochastic model for machine learning mixed memberships. Airoldi et al. (2005) have adapted the model to examine single-mode network data, yielding novel clusterings in protein–protein interaction networks (Airoldi et al., 2006). The primary distinction between FOG and relational LDA models is that LDA allows a single observed to be explained by a mixture of groups, whereas FOG assumes that a single social context is associated with a given observation, but hierarchically clusters such contexts to construct a restricted mixed-group structure.

2.3. Link data from network data

Link analysis and network analysis have grown out of distinct communities, despite being frequently applied to the examination of the same interaction phenomena. In many ways, grouping research has become an intersection point in which practitioners of both fields are attempting to capitalize on the strengths of the other. Link analysis researchers approach group models as an opportunity to characterize structure and dependence in interaction data which is too often analyzed as though observations were independent. Analysts who have traditionally used graph theoretic approaches to examine network data are incorporating statistical models and significance tests to improve their ability to reason about noisy data and support claims about the significance of structural characteristics in their networks. For frameworks such as FOG to see the widest use (and scrutiny), we must develop translation techniques that allow data in one format to be examined using algorithms for the other. These translations must account for disparate data qualities emphasized in the two branches of analysis.

Since small changes in network structure can have a large impact on the graph theoretic measures used in network analysis, translations from link data to network data are designed to reduce noise as much as possible. Many network datasets begin life as something more closely resembling link data. Lists of interactions or survey responses are “flattened” into a matrix of pairwise interactions using summation, cutoffs, or reciprocation criteria depending on the interaction being studied and the network type desired. Recently, Kubica et al. (2003a) have presented cGraph, an expectation maximization approach to detecting underlying networks.

The stochastic link analytic techniques we examined are intended to robustly handle noise given enough data. However, when our source data is limited to the information in an interaction matrix, we run the risk of amplifying any noise present when we generate additional links. We must also tackle the problem of reflecting the structural data contained in the network model in a way that link analytic algorithms can interpret.

The simplest approach would be to interpret each edge in the network data as a single link between two entities. Though it retains all of the original data, this naïve method produces links that individually contain the minimal amount of structural information. Broader patterns such as paths and clusters can be revealed only by inspecting many links at once. This disadvantages greedy algorithms that examine individual links, such as H-FOG, because they are provided little basis on which to make their earliest (and most important) clustering decisions. For these algorithms, we must generate “richer” links that give more structural information while still only sampling the overall network.

In this paper we have adopted the “random tree” solution described by Kubica et al. (2003a), inverting its purpose to generate random interactions rather than extract graphs from observed interactions. Link data is constructed stochastically by iteratively adding to links entities which are randomly chosen from the neighbors of those already present. Fig. 1 illustrates this process. In the illustration, nodes A and B have been visited already, making the entire peripheral boundary of C, D, and E available as possible next additions. Note that C has a twice greater probability of selection than D or E, as there are two links proceeding to it from our already-visited structure. If our matrix were weighted, it would be the summed weights of those links rather than the quantity that determined relative likelihood. The intuition behind this process is to simulate chains of gossip or casual assemblies that would be directly measured as a source of 2-mode data if available. We use this technique on a weighted collation of Sampson’s monastery data later in this paper.[1]

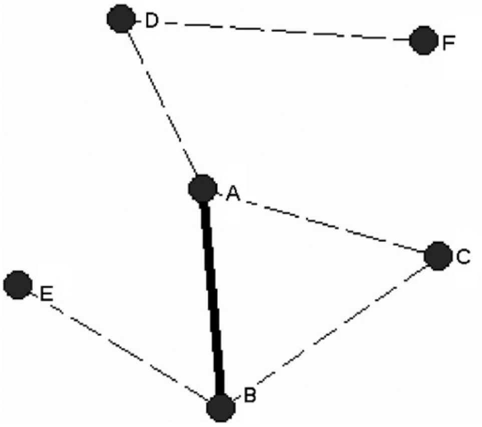

Fig. 1. Random tree in progress. (A) and (B) have been visited; (C), (D) and (E) are candidates for next added node.

Prior work has examined relationships between networks and distributions of random processes on them. Kashima and Tsuboi (2004) showed that random walks can be used as a kernel in classification of structural features of a graph. Random walks and trees in social networks have been used in simulations as analogs to real-world processes, such as knowledge dissemination or the spread of a disease (Christley et al., 2005). Page and Brin (1998) note that eigenvalue centrality of each node in a network is proportionate to the fraction of an infinite length random walk it will occupy, which they have famously analogized to the search for information on the web.

In this paper, we interpret groups found in random trees as sets of individuals who are likely to be exposed to the same experiences. Since we cannot directly calculate or represent the distribution of such groups, we sample them instead by computing a fixed number of groups of fixed size.

2.4. Datasets

2.4.1. Sampson monastery

We chose Sampson’s monastery dataset (1968) as a testbed for the FOG framework because it is one of the datasets most widely discussed in social grouping literature. Sampson conducted a survey in which novice monks at a monastery ranked their compatriots according to four criteria: like/dislike, esteem, personal influence, and consistency with the creed of the monastic order. Sampson made strong arguments for several discrete social groups in the data based on direct anthropological observation. Events confirmed his observations when, during the study, novices of one group resigned or were expelled over religious differences. Samson’s surveys may be the dataset that comes closest to providing social data with a labeled “ground truth” for grouping research.

[1] A potential criticism of this method is that its stochasticity introduces uncertainty into results. In fact, since results will converge with a large sample, the user can define a preference between accuracy and speed by specifying a sample size. Reproducibility can be achieved by storing and reusing a random seed. Finally, random link production is consistent with FOG’s modeling of uncertainty in all data, and we believe affirmed by both our own empirical results and the prior efforts below.

Sampson's monastery is discussed in greater depth in Sampson's original (1969) dissertation, and in the December 1988 issue of the journal *Social Networks*. We compare the groups discovered by FOG to Sampson's and those presented by Reitz in that issue in his introduction of a hierarchical clustering algorithm. Like that paper, we use Breiger et al.'s (1975) collation of Sampson's data: for each of the relations "like", "esteem", "influence", and "consistency", the top three positive selections by each individual at time three are recorded in a relation matrix. Negative selections are ignored, as negative relations are intransitive and thus cannot be positive evidence of an inherently transitive co-membership.[2] These matrices are summed, yielding a single matrix summarizing the preferential data at that time period. The matrix in its entirety is shown below as Table 1. Because FOG analyzes link-based data, we then pre-process this matrix to generate links using the random tree technique described above.

2.4.2. *Davis, Gardner, and Gardner's southern women*

The southern women dataset (Davis et al., 1941) lists the attendance of 18 women and 14 parties. The parties in this network are precisely the type of linking observation which FOG is designed to analyze without pre-processing. As with the monastery dataset, there exists a labeling for groups based on direct observation rather than algorithmic analyses. Davis et al. used ethnographic analysis, including surveys, to distinguish not only between the two major cliques, but three tiers of centrality within them.

A wide variety of mathematical approaches have been used to reanalyze the data. Freeman (1992) preformed a comprehensive meta-analysis of 21 such studies, and we analyze our results in response to some of his conclusions. We have also accepted that paper's verdict on which of two conflicting figures in the original work was correct. We reproduce that figure as Table 2, below, for reference.

3. The FOG framework

Grouping methodologies are often introduced as algorithms, although they encompass distinct models, measures, data translations, and validation schemes as well as the model-fitting algorithm itself. To minimize this confusion, we discuss FOG as a framework consisting of several components. The FOG generative model relates link interactions we observe to group entities, which are hidden. The H-FOG algorithm is a simple link-clustering approach to fitting groups of the type described in the model to data. (As we will show, the algorithm does not guarantee optimality and future work may yield a fast algorithm that finds better fits.) A separate link generation algorithm creates link data from social network data.

3.1. *Stochastic model of evidence generation*

Since we are trying to infer groups based on link evidence, we define our group membership relation as the tendency to be produced in observations associated with the group. We can alter the strength of the tendency to be included in observations without altering its fundamental character. We formulate the above mathematically as follows.

Table 1
Breiger et al. (1975) collation of Sampson survey data

		ROMUL	BONAVEN	AMBROSE	BERTH	PETER	LOUIS	VICTOR	WINF	JOHN	GREG	HUGH	BONI	MARK	ALBERT	AMAND	BASIL	ELIAS	SIMP
		10	5	9	6	4	11	8	12	1	2	14	15	7	16	13	3	17	18
ROMUL	10	0	1	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
BONAVEN	5	0	0	2	0	3	3	0	0	0	1	0	0	0	0	0	0	0	0
AMBROSE	9	0	1	0	0	2	0	2	1	2	1	0	0	0	0	0	0	0	0
BERTH	6	0	1	3	0	4	2	0	0	0	1	0	0	0	0	0	0	0	0
PETER	4	3	1	0	4	0	4	0	0	0	0	0	0	0	0	0	0	0	0
LOUIS	11	0	3	2	0	2	0	2	0	0	0	1	0	0	1	0	0	0	0
VICTOR	8	0	1	2	3	4	2	0	0	1	0	0	0	0	0	0	0	0	0
WINF	12	0	0	0	0	0	0	0	0	3	3	1	0	2	0	0	0	0	0
JOHN	1	0	1	0	0	0	0	1	4	0	1	2	0	1	0	0	1	1	0
GREG	2	0	1	0	0	0	0	1	3	4	0	0	0	3	0	0	0	0	0
HUGH	14	0	0	0	0	0	0	0	3	4	3	0	4	0	1	0	0	0	0
BONI	15	0	0	0	0	0	0	0	1	2	4	3	0	2	0	0	0	0	0
MARK	7	0	0	0	0	0	0	0	3	0	4	0	2	0	4	0	0	0	0
ALBERT	16	0	0	0	0	0	0	0	1	0	4	0	4	4	0	0	0	0	0
AMAND	13	0	4	0	0	0	3	0	0	0	0	0	0	3	0	0	0	0	1
BASIL	3	0	0	0	0	0	0	0	0	4	0	0	0	0	0	4	0	4	2
ELIAS	17	0	0	0	0	0	0	0	0	0	3	0	0	0	0	1	3	0	3
SIMP	18	0	0	0	0	0	0	0	0	1	4	0	0	0	0	0	3	4	0

[2] FOG is agnostic regarding the definition of "association" within the group it detects; instead the definition is defined implicitly by the input. For example, enmities and friendship ties might be provided as equally compelling evidence of competitive association. Alternatively, to detect only internally friendly groups, one would provide FOG with only friendly ties. In Section 4, we compare to a contrasting approach by Doreian and Mrvar (1996) which incorporates evidence of active disassociation.

Table 2
Southern women party attendance, reproduced from (Davis et al., 1941)

Names of participants of group I		Code numbers and dates of social events reported in *Old City Herald*													
		(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 3/23	(12) 4/7	(13) 11/21	(14) 8/3
1.	Mrs. Evelyn Jefferson	×	×	×	×	×	×		×	×					
2.	Miss Laura Mandeville	×	×	×		×	×	×	×						
3.	Miss Theresa Anderson		×	×	×	×	×	×	×	×					
4.	Miss Brenda Rogers	×		×	×	×	×	×	×						
5.	Miss Charlotte McDowd			×	×	×		×							
6.	Miss Frances Anderson			×		×	×		×						
7.	Miss Eleanor Nye					×	×	×	×						
8.	Miss Pearl Oglethorpe						×		×	×					
9.	Miss Ruth DeSand					×		×	×	×					
10.	Miss Verne Sanderson							×	×	×			×		
11.	Miss Myra Liddell								×	×	×		×		
12.	Miss Katherine Rogers								×	×	×		×	×	×
13.	Mrs. Sylvia Avondale							×	×	×	×		×	×	×
14.	Mrs. Nora Fayette						×	×		×	×	×	×	×	×
15.	Mrs. Helen Lloyd							×	×		×	×	×		
16.	Mrs. Dorothy Murchison								×	×					
17.	Mrs. Olivia Carleton									×		×			
18.	Mrs. Flora Price									×		×			

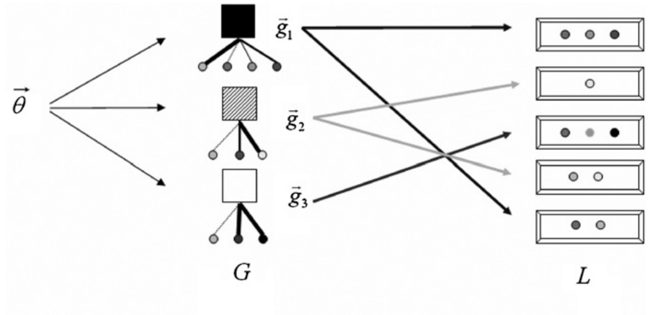

Fig. 2. Illustrating relationships in the FOG model.

Consider a set of entities E and a set of groups G. Entities are elementary objects whose presence or absence is observable in a set L of links (which are sets of entities). Groups emit pieces of evidence, which consist of sets of entities which are co-observed. Groups emit with different frequencies, according to a probability distribution $\vec{\theta}$ across groups such that θ_g, for $g \in G$, is the probability that any given link was emitted by group g; $\sum_{g \in G} \theta_g = 1$. Elsewhere in this paper we refer to θ_g as the *emission prior*, since it represents our expectation that a piece of evidence will come from a specific group, prior to examining the members observed in the link. A membership vector $\vec{g}$, whose entries are the probabilities that that each entity is present in a link that has been emitted by group g, further describes each group. We write this as $g_e = P(e \in l | g \Rightarrow l))$, or the shorthand $P(e|g)$. We will refer to g_e as membership strength or affiliation. Fig. 2 illustrates the hierarchy of objects we have defined.

When considering the likelihood that a particular group would produce a specific link, we must consider not only the probability of observing the entities present in the link but also the probability of excluding those not present.

$$P(l|g \Rightarrow l) = \left(\prod_{e \in l} g_e\right)\left(\prod_{e \notin l} 1 - g_e\right) \tag{1}$$

The assumption that, in the emissions of a single group, members are emitted completely independently is important to maintaining that the membership relation differs only in intensity between entities. (A joint distribution would imply additional substructure.) Similarly, we assume that links are generated completely independently given the groups and their emission priors, so that the only structure exists between the groups and the entities themselves, and in the relative frequency of emission of the groups. Combining, these we can derive the likelihood that an entire set of evidence would be produced given a grouping and an emission distribution vector. The factorial coefficient in this equation normalizes for the ordering of the link set, which is irrelevant to our model.

$$P(L|G, \vec{\theta}) = |L|! \prod_{l \in L} \sum_{g \in G} \theta_g \left(\prod_{e \in l} g_e\right)\left(\prod_{e \in l} 1 - g_e\right) \tag{2}$$

Performance and representation precision (probabilities involved can be extremely small) demand that the above likelihood function be calculated via a log-likelihood transformation. To enable this transformation, we place the restriction that $g_e \in [p_{\min}, p_{\max}]$, where $0 < p_{\min} < p_{\max} < 1$. This ensures that a group always has some nonzero probability of emitting its least related entity, or excluding from a link even its most significant member.

Previous stochastic models of link generation have included an “error term” under which there is some small probability that a link will be emanated containing entities which do not cohabit any

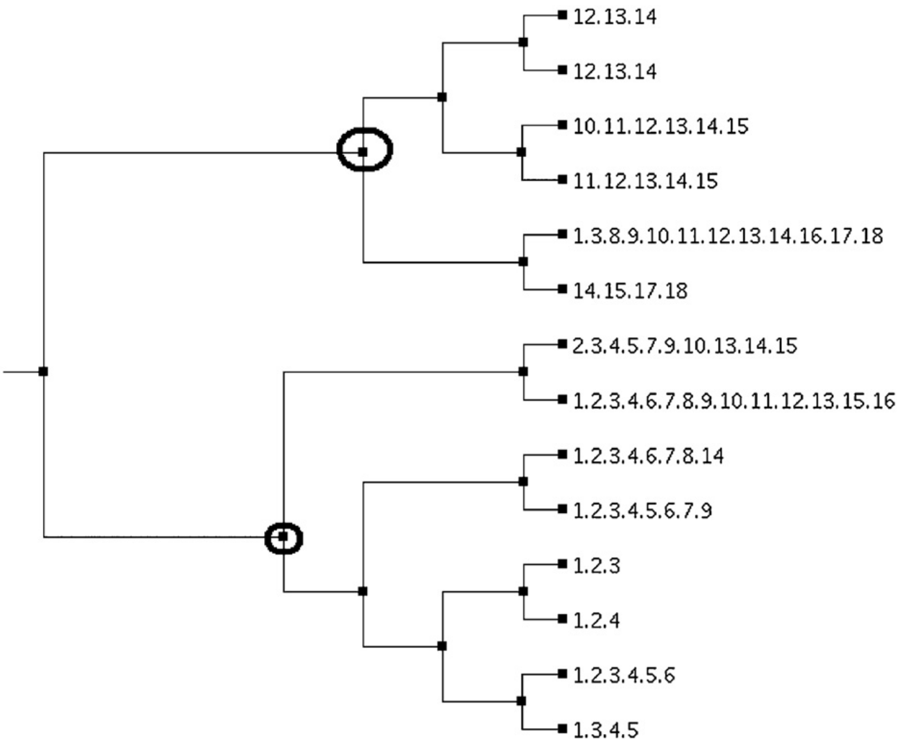

Fig. 3. Clustering tree from the DGG dataset.

group. This was necessary to allow models to be fit to data without placing extreme penalties on groups which were forced to include outlying entities as equal members to more supported nodes. In FOG, a similar purpose is served by allowing weak memberships and assuming weak universal memberships, with the advantage that we need no prior beliefs about an error rate.

3.2. The H-FOG algorithm

With the relationship between groups and evidence described above, we can reduce group detection to an optimization problem which searches for the grouping with the greatest likelihood of generating the observed evidence: $\arg \max_{G,\vec{\theta}} P(L|G, \vec{\theta})$. Calculating this explicitly would be intractable, so we propose an estimation algorithm.

Since our model requires that a single group be responsible for emanating each link, we can restrict our search by considering only groups which optimally represent some partition of the data. The group with the highest probability of single-handedly generating a set of links is the one which emits each entity with probability equal to the proportion of the link set in which that entity occurs. We build groups of this sort by iteratively clustering link evidence in a way that ensures links with the greatest similarity are grouped together. For each pair of groups g_1, g_2, we consider a new group g_n that would maximize probability of emitting the combined evidence supporting both groups ($L_n \leftarrow L_1 \cup L_2$). We then calculate the ratio as a heuristic indicating the relative increase in likelihood of the underlying links if they are considered the emissions of one merged group rather than two separate ones.

$$\frac{|L_n|^2 P(L_n|g_n)}{|L_1|^2 P(L_1|g_1) + |L_2|^2 P(L_2|g_2)} \qquad (3)$$

The pair for which this ratio is highest is merged.

The tree in Fig. 3, constructed from the southern women dataset, illustrates the hierarchical clustering of evidence. Each intermediate node corresponds to a group tuned to produce evidence of the types found in the leaves below. We define a *horizon* from this tree as a set of nodes such whose children span all of the evidence, for which none is the ancestor of another. A horizon, such as the circled nodes in Fig. 3, corresponds to a set of groups which account collectively for all of the observed evidence. If we choose our horizon from the bottom level, groups are tuned to very specific profiles of evidence, so that they are expected to produce any of the few links below them with relatively high probability. As we move up the tree, membership rosters for groups become more complex and the distribution of links which they produce becomes more entropic, so that the probability of producing any particular link drops exponentially. At the same time, θ_g values rise as we ascend the tree, since each group represents a greater proportion of the underlying evidence. Near the top, groups are overly general and fit the evidence underneath poorly, so that, even though θ_g's are high, the total probability of producing the evidence set is quite low.

Unfortunately, reduced $P(l|g \Rightarrow l)$ outpaces increased $P(g \Rightarrow l)$ over a climb of the tree, so that there is usually no optimal midpoint that would allow us to discern a “most probable number of group entities”. This conflict between the need for well-supported groups and ones that tightly fit the data must be resolved by a preference fitting the context of the analysis. As such, an operator must currently specify a number of groups, k, for which to search, effectively deciding on a tolerable tradeoff between a simple model with few groups and a model which most closely fits the evidence but may in fact be over-fit. Fortunately, due to its hierarchical nature, H-FOG needs be run only once to generate candidate groupings for each feasible k. An analyst can then explore different numbers of groups dynamically to determine subjectively which is best supported.

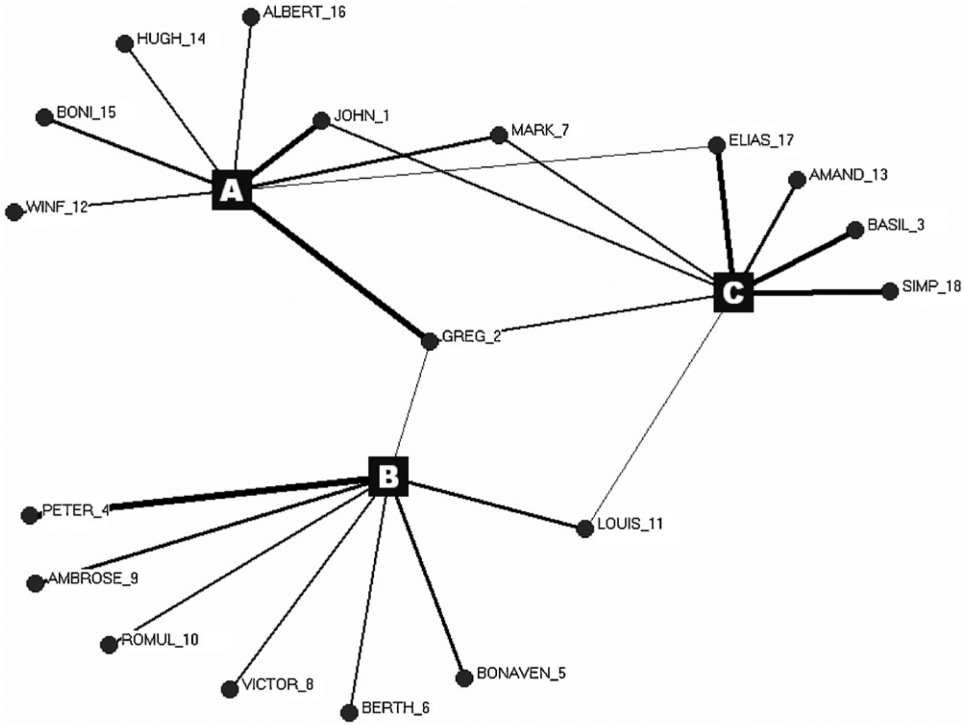

Fig. 4. Fuzzy groups in Sampson's monastery. Group A corresponds to the "young Turks", group B to the "loyal opposition", and group C to the "outcasts".

4. Results and analysis

4.1. Fuzzy groups in the monastery

Since Sampson's data consists of pairwise relations, we generated link data using the random-tree technique previously discussed. Ten trees were initiated at each node, each expanding to contain three nodes, and the set was clustered using the FOG algorithm. Results are shown as a two mode (agent → group) network in Fig. 4. Line thickness indicates the degree of membership. We normalized line thickness within each cluster because it is not necessarily appropriate to compare association levels between groups. This is because our link generation method required exactly three individuals in each observation, artificially deflating the average frequencies of emission in large groups and inflating it in small ones. Nodes have been manually laid out to elucidate membership categories we discuss.

Sampson identified novice 2, Gregory, as the most significant leader of the "young Turks", the liberal newcomers who would be expelled or resign in the coming drama. The members of that group are collected exactly as those affiliated with group A. Gregory's position, as both the most affiliated to the Turks and the only novice with connections to all three groups, suggests a high degree of centrality both within the young Turks and in the network as a whole. This type of border-spanning centrality has been linked to iconoclasticism, power, and stress, painting a vivid picture of factors which may have contributed to Gregory's exit. The official reasons given for his expulsion were excessive independence and arrogance. Could he have been singled out as more dangerous precisely because he had captured the attention and esteem of individuals outside the clique? Rank of affiliation to group A turns out to be a good predictor for the order in which the young Turks left the abbey: the second individual most affiliated with the Turks is Sampson's second identified leader, John (novice 1), who resigned shortly after Gregory's exit, trailed by Mark (7) and the other novices.

Sampson's third and lowest-ranking leader, Winfred (12), does not buck this trend. Although his relatively low association to group A belies his eventual identification as a leader of the Turks, it accurately reflects his placement as the last one to leave the Abbey. Winfred's undistinguished position in our plot illustrates some biases of this analytic method, as well as some peculiarities of his situation. Winfred identified strongly enough with his group that he was completely embedded: all of his incoming and outgoing connections in the survey data lie within the Turks. The result is that, although the random trees in which he appears are exclusively tied to group A, he simply does not participate in nearly as many total evidence pieces as high-betweenness boundary spanners like Gregory or John. As this shows, our evidence-generation technique could rightly be said to put a premium on individuals with high betweenness. However, this is defensible when paired with interpretation placing it in a social context. Recall that we generated random trees in order to model iterative interaction processes within the graph, such as the spread of a rumor or the slow accumulation of individuals to a casual gathering. It is easy to imagine an individual with more diverse ties, such as Gregory, being drawn into a wider variety of gatherings. By being a prolific interactor, Gregory may well have defined the Turks to the rest of the community, without necessarily intending to or even identifying exclusively with them.

Evidence supports the distinction between Gregory's "celebrity" and Winfred's "poster child" stances. In Sampson's study, Winfred's leadership was either absent or unobserved in the presence of the two higher-profile leaders, and became clear only after their exit. Winfred's embeddedness seems to reduce his significance at the time of our analysis, but as the split widened between the Turks and the opposition, making positions like Gregory and John's untenable, Winfred's exclusive loyalty became the crucial element of his in-group leadership.

The membership and leadership of the "loyal opposition" party are similarly gathered around group B in our plot. Peter (4) and

Bonaven (5), who were identified by Sampson as the leaders of the opposition, show the highest affinity for the group. Members described as less attached show less affinity, and one such novice shows a split allegiance to the outcast group.

The absence of any links, save Gregory, between the opposition and the Turks serves to reflect the conflict between the two groups. By contrast, the "outcasts" in group C have several members associated with other groups as well. These cases show that fuzzy memberships can help elucidate not only the complexity of an individual's allegiances, but also the character of a group as exclusive or inclusive to interstitial members.

Sampson originally identified a fourth group, but we restricted our analysis to three clusters because the last was not a cohesive group fitting our definition. Sampson does not describe the "waverers" as a set of individuals allied or interacting with one another, but as being in similar positions of doubt between the two major groups, more akin to our interstitial roles. Additionally, previous analyses have questioned the distinction between the waverers and the loyal opposition. Our own analysis places two of them, Romul (10) and Victor (8), as weak members of the loyal opposition. From a purely structural perspective they are tied more to the loyal opposition; whatever there mental allegiance.

Armand (13) is categorized as an outcast, owing less to his statements of affinity for those individuals than from Basil's (3) and Elias' (17) connections to him. Our classification of Armand as an outcast is in line with the discrete partitioning provided by Doreian and Mrvar (1996), who demonstrate that there was increasing evidence over time that this foursome was a genuine group. Doreian and Mrvar used a block modeling approach optimizing structural balance, a measure of cohesion incorporating both positive and negative relations. Interestingly, their partitioning is perfectly correlated with the groups to which each individual is assigned maximal membership by FOG. We take these convergent results from different methodologies as encouraging validation in a setting for there is no known ground truth. The Doreian and Mrvar study also includes a temporal analysis suggesting that in the final period of Sampson's observation, two members of the Young Turks, Gregory and Mark, gave responses that fit better within the Outcasts partition. Both of these individuals are marked as interstitial in the FOG results, although Gregory's departure is somewhat surprising considering his very strong alignment with the Young Turks and weak connection to the Outcasts.

4.2. *Fuzzy groups among southern women*

Analyses of the DGG data, including the original, have generally partitioned the women into two cliques[3] that intersect on a few individuals or events. We use a "spectrographic[4]" visualization scheme in Fig. 5 to present the results of a 2-clustering of the southern women in greater detail than would have been readable in the Sampson analysis. Bars of each color indicate each woman's affiliation with two groups derived from 8 and 6 of the party rosters respectively. Individuals are sorted along the *X*-axis according to the difference in their membership levels, which maximizes the visual distance between the cliques. We have also included a 2-mode network visualization for comparison to the one we presented for the Sampson data.

The results of our algorithmic approach correspond strongly to the intuitive conclusions of Davis et al. In group A, the core and primary periphery are reproduced precisely as plateaus in the membership levels. Someone attempting to fit our analysis to their mode might draw slightly different tiers for the group B, but the rough ordering of individual affiliations is the same. For both groups, the most peripheral members are seen in the center of our chart, with low levels of affiliation in both groups. Some of these members have been shown to be interstitial; for example Davis et al. report that Ruth (9) was claimed by both cliques in interviews with members. Others, such as Pearl (8) and Verne (10) were only claimed by members of the cliques to which our chart shows greater affiliation.

There are many mathematical studies of the DGG data to which the H-FOG clustering correlates. We will omit a pairwise comparison, as many of the results are significantly similar to Davis et al.'s intuitive analysis described above, and a comprehensive meta-analysis has already been accomplished by Freeman. Instead, we focus on FOG's contribution to one prong of that analysis: the core-periphery structure of the two cliques.

Davis et al.' describe as "core" the individuals that are seldom excluded from their clique's functions. We see that the most affiliated individuals in both groups demonstrate a propensity to appear with the other group as well. This supports the argument we proposed with the Sampson data,[5] that leaders of a group may either arise out of greater participation with other groups than do the less active members, such as those in DGG's "primary" and "secondary" members, or else experience more pressure to do so.

In his meta-analysis, Freeman treated core-periphery as an ordering of individuals for each group, without specifying that centrality in one group promoted distance from the other (although that was a side effect of many techniques compared). FOG results certainly fit that mode, but the juxtaposition of affiliations given above lends itself to an additional breakdown of several interactions. We can separate individuals into several modes of interaction. We have central leaders, such as the novices John or Gregory or the Southern women Nora and Katherine. There are embedded leaders such as Winfred, Laura, or Brenna. There is a loyal second tier in each of the groups we have analyzed, and finally a set of truly interstitial individuals who participate at low levels in both groups.

From our observations of these roles in Sampson data, we might issue the prediction that a thoroughly embedded member, such as Brenda or Flora, would flourish if there were a falling out between the two groups. On the other hand, if good relations continued between the groups, our profile of an emergent leader might better fit individuals such as Ruth or Helen: those with strong ties to one group, but some degree of participation with the other. Davis et al. do not examine conflict between these cliques and describe no events that would be telling regarding our first hypothesis. However, they completed a larger study of many cliques, in which they used interstitial members to examine relations between social classes associated with each clique. They describe a class of "on the way up" individuals, who participate in events outside their clique in order to socialize with those above them in social class.

The interpretation of interstitial members as a separate class is supported by Doreian et al. (2005), wherein an error-averse block-model partitioning of events revealed that Pearl and Dorothy attended *only* events attended by members of both groups. For the block modeling approach, which considers extra-group connections when assigning groups, this was sufficient evidence to place them in a distinct group. FOG, which permits overlap but optimizes only for intra-group cohesion, places them instead in both groups. By analyzing interstitial members, we can use FOG to capture some

[3] We use *clique* here to maintain consistency with prior work, not to indicate a graph theoretic relationship.

[4] So named after similarity to overlaid graphs of element density used to differentiate substances in mass spectrometry.

[5] Since the DGG analysis is based on direct observations rather than synthetic observations from random trees, we do not have the same concern about overemphasizing centrality that we did with the novices.

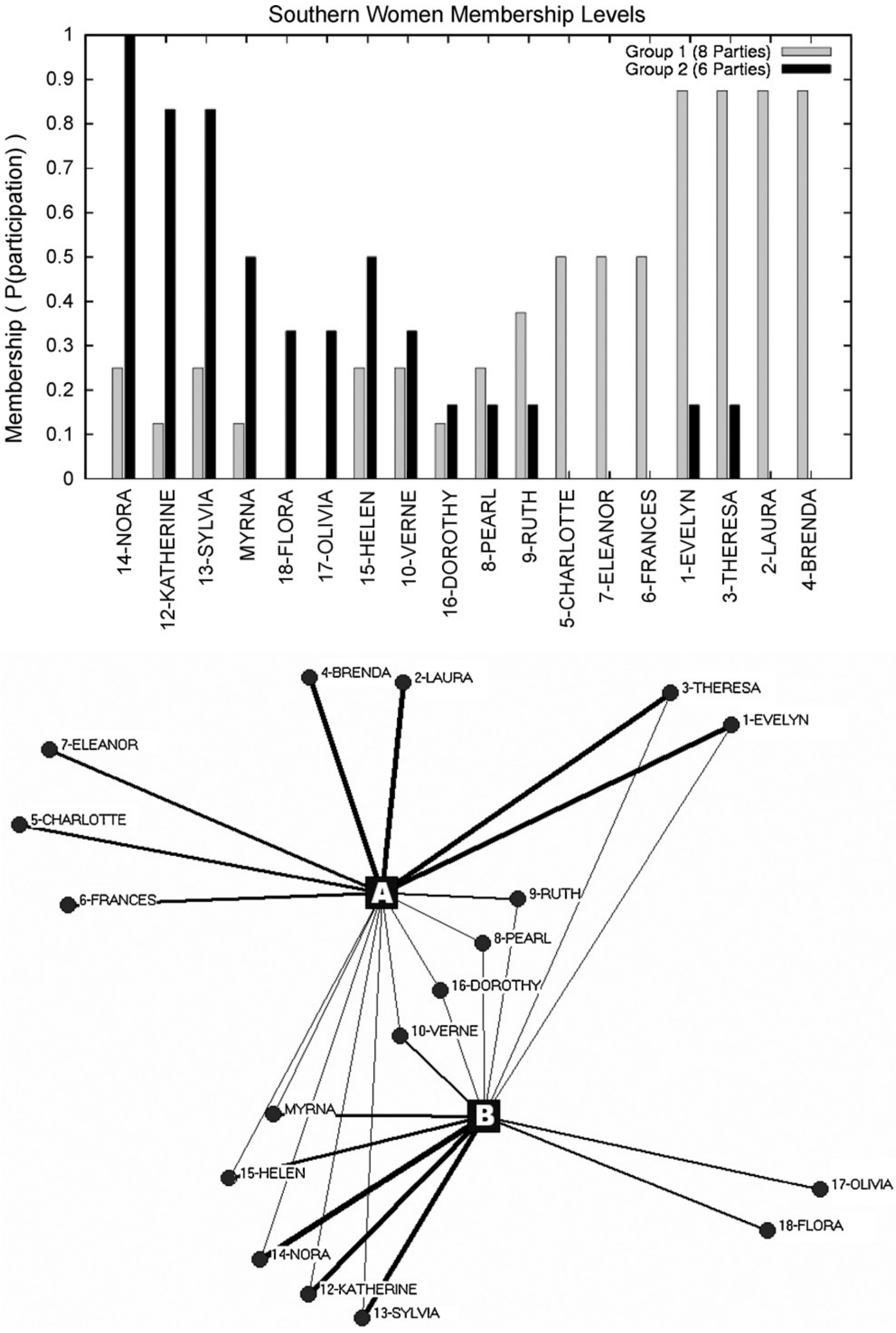

Fig. 5. H-FOG 2-clustering of the DGG dataset, spectrographic (top) and network (bottom) representations.

of the same structural insights provided by block modeling. However, some structural subtleties would not be captured by FOG. For example, we would be unable to distinguish between individuals like Pearl and Dorothy that attended only mixed events, and individuals who attended no mixed events but attended some from each group.

4.3. Runtime

The graph in Fig. 6 profiles the runtime of the H-FOG algorithm on evidence generated from random groupings with varying number of entities. Memberships for each entity are chosen from a uniform distribution, and several sets of evidence with varying numbers of links are generated, according to the FOG stochastic model described previously. In total, the figure summarizes runtime on 1500 evidence sets from 150 groups. The algorithm was implemented in Python, an interpreted scripting language, and executed on an Intel Pentium 4 machine running at a speed of 3 GHz.

Number of groups (not shown) has a minimal effect on runtime, as the vast majority of calculation is performed on the lower levels of the tree, before the cutoff for number of groups is reached.

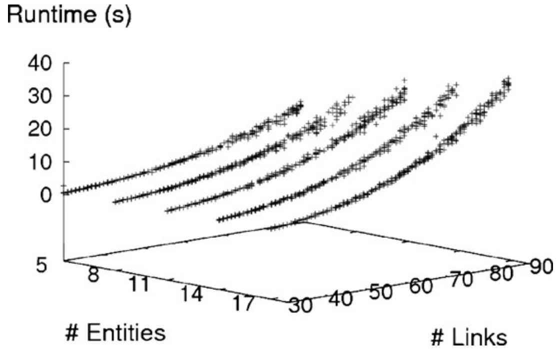

Fig. 6. FOG Runtime v. # entities and # groups.

The effect of the number of links (L) being grouped dominates that of any other variable on runtime. Runtimes of H-FOG fall under an O(L^3) bounding. This is expected from the fact that H-FOG must examine O(L^2) candidate merges at each of L levels in the merge tree. Runtime is affected linearly by the number of entities observed, as this number determines the upper bounds on the number of calculations necessary to examine a candidate group. In practice, many of these calculations are memoized, allowing for a tighter bound than discussed.

5. Discussion and future work

We set out to introduce a new quantitative way of reasoning about the complex relationship between individuals and groups, allowing varied degrees of participation in multiple groups. We proposed the FOG stochastic model, which dictates relationships between individuals, groups, and observable interactions as a generative model for link data. To make FOG a useful analysis tool, we introduced the H-FOG algorithm which fits a model to existing link data. To investigate single-mode network data, we implemented a simple method for generating rich multi-entity links from a pairwise network based on a simplistic simulation of interaction processes.

5.1. Validation

Mathematical approaches to group detection are based on the assumption that groups have a reality outside individual perceptions, which we can detect statistically. It should therefore be possible to empirically validate grouping methods on their ability to predict the outcome of processes in which groups play a role. Currently, the closest we have to this sort of predictive test is to compare our analysis to that of anthropologists like Sampson, who were able to relate their intuitive observations to unforeseen events in the social group. Can fuzzy grouping rediscover social patterns that stood out to ethnographers in the field?

In the two datasets we studied, the answer is yes. The discrete groups identified by both Sampson and the DGG team were nearly identical to the list of individuals with greatest affiliation to each group in our analysis. Additionally, substructures and leadership roles identified by the original authors corresponded strongly to the levels of affiliation we discovered. FOG sits well among a variety of mathematical approaches which have supported the original intuitive analyses. However, these have usually relied on separate techniques to distinguish groups, leaders, and internal structures. One advantage of FOG is the ability to unify these multiple levels of analysis under a simple model.

On the subject of validation, many link analytic methods, including k-Groups and iterative deduplication, have been validated from a data mining perspective by testing the ability of the method to rediscover groups from artificially generated data. We plan to conduct this type of examination when we complete a new fitting algorithm to replace H-FOG.

5.2. Interstitial roles

The existence of interstitial roles, where an individual retains several group affiliations, was our principle motivation for developing a fuzzy grouper. We identified many such individuals in our analysis, fitting several profiles. With great frequency, the most apparent leaders of a group had weak ties to other groups as well, as did those members with the least affiliation to any group. The differentiation of these two roles, as well as the surprising result that most groups contained a well-embedded middle tier, would be difficult without FOG's novel properties: the combination of multiple memberships and degrees of membership. As FOG is applied to additional datasets we expect that a better understanding of individual roles based on multiple memberships will emerge. The FOG approach holds promise of providing a mathematical base for capturing and defining some critical types of social roles not heretofore measurable.

It is worth noting that FOG did not always identify as interstitial the individuals whom we would expect. In some cases, such as with Sampson's waverers, individuals who were considered interstitial by an observer were placed in single groups by FOG. Conversely, some of the "secondary" clique members in the DGG dataset would appear to be interstitial on a reading of our charts, but were only claimed by a single clique in specific surveys conducted by Davis et al. The distinction between members who are simply weakly connected and those who fill an actively interstitial role may be beyond our level of analysis. Alternatively, noise may have been introduced in the specific data we examined, or results may have been misinterpreted by the original observers. Since analysis of interstitial roles is a vital component of FOG, future work should investigate in depth what factors in data affect our ability to differentiate roles.

5.3. Generating link data from networks

Although the theory underlying the FOG model requires link data that indicates a shared context between members, we are optimistic about the ability to examine single-mode network data by generating fake data from simulated interactions. In the Sampson data, we were able to affirm existing knowledge about the monastery social groups using this approach, while generating new theories.

A crucial aspect of this analysis was to connect the final results with the assumptions under which link data was generated. Since Breiger et al.'s matrix indicated relationships between novices that could lead to interaction, we built our link generator as a simulator of social contexts that spread "infectively" through iterative interactions. This type of link increased the observation frequency of high-betweenness individuals, but we might expect those individuals to be disproportionately represented in real data recording this type of interaction. Understanding this bias helped us interpret the difference between embedded and interstitial members when interpreting the role of novice Winfred (12), the last leader of the young Turks.

A potential criticism of random link generation is that it injects variance into our analysis. In this study, we approached the problem by generating larger sets of random trees until differences between runs were below the threshold of our qualitative analysis. However, increasing the number of samples comes at a significant computational cost. We attempted a similar process using a network from a cleaned corporate email corpus (Diesner et al., 2006) containing 150 users, and were still experiencing visible variance between results when using samples of 450 links at an average runtime of over 8 h. We hope to address this in future work, first by improving scalability so that larger samples can be used (see *alternative algorithms* below), and second by conducting a sensitivity analysis to guide selection of sample parameters for networks of different sizes and properties. Notably, some networks drawn from 2-mode data may be easier to analyze in their original form. We were able to perform an informative analysis of the same emails in only 30 min by multi-recipient emails rather than random walks on the incidence network as observations.

Another peculiarity of the random-tree link model is that it discards the directionality of links. Since FOG interprets only the presence or absence of an individual in a link, no distinction is drawn between individuals originating a random observation and those added subsequently. This affects the placement of individuals like Amand (13), who appeared in many interactions with individuals whose admiration or affection he did not reciprocate. One could again argue that many types of real data would have similar confusion, but it is also possible that a link model could be adjusted to include this information.

We also believe that networks of different relations may require different link models. In a formal communication network, such as a corporate hierarchy, where messages pass along a fixed route from source to destination, a random, directed walk would be more appropriate than a random tree. It might be convenient to analyze 2-mode networks by simply interpreting one of the modes as links, but that decision should similarly depend on the type of relationship represented in the network. Link generation for multi-mode networks is another direction we intend to develop to further FOG's applicability.

5.4. Analyzing and visualizing fuzzy relationships

Social groups with binary memberships can be analyzed by common statistical techniques. For example, when Davis et al. introduced the southern women dataset as overlapping cliques, they were able to investigate the character of each clique by taking aggregate statistics over its members. The same analysis would be non-trivial for a FOG cluster. What is the mean income of the members of a fuzzy group? The question is especially difficult because our results are intended to denote a level of participation, and not necessarily the degree to which members are representative of their group. If fuzzy groupings become a useful analytic tool, new measurements will have to be developed or adapted to properly interpret the new information given.

We have barely scratched the surface on that work in our intuitive analysis of the clusterings in this paper, but we have tried to uphold several principles in our analysis. First, membership values should not be examined independently of the context of other memberships held by the same individual and to the same group. Groups or individuals may have different average memberships, for reasons that have less to do with the actual importance of those memberships than with the nature of group events or the way data was collected or generated. Secondly, the novel strength of grouping with multiple, variable memberships is the ability to compare several simultaneously occurring memberships in individuals. We intend use FOG to define and investigate roles that are defined in terms of multiple memberships, rather than to rehash issues of internal group structure that have been examined by other algorithms.

At this phase in our understanding of fuzzy overlapping groups, visualizations play an especially important role by influencing the types of patterns we can identify intuitively. We have presented two visualization paradigms in this paper, one indicating individuals' memberships to groups as a weighted two-mode network, and the other a spectrographic view providing all membership levels explicitly in bar chart form. As with most visualizations of overlapping clusters, placement of individuals can be difficult as the page does not have enough dimensions to represent all association patterns. We had few enough groups in both of our analyses that we were able position individuals for reasonable clarity, but this would not be true in more complicated datasets. We have experimented with several heuristics for laying out more than two groups in spectrographic figures, but more work needs to be done in this area.

5.5. Alternative algorithms

In this paper, we focused on one algorithm, H-FOG, which performs hierarchical clustering on event data. H-FOG was appropriate for the datasets we examined because of their relatively small size and known number of groups. We are currently investigating alternative algorithms which share the FOG stochastic model but allow settings with more data or unknown group numbers.

If an analyst plans to examine only a specific number of groups k, then H-FOG waists significant computation time calculating branches of the tree above and below that point. k-FOG is an alternative algorithm which uses expectation maximization (EM) clustering rather than hierarchical. Like H-FOG it is susceptible to local maxima, but tends to converge much faster and is thus suitable to larger datasets. In some settings, an analyst may be interest in groups of a certain type but be uncertain of the exact number. α-FOG allows specification of an α parameter encoding the desired level of group cohesion (as compared to number of groups). It then approximates optimal group assignments and number of groups using a Dirichlet process built around the likelihood function described in this paper.

The details of these algorithms are beyond the scope of this paper, but future work comparing them will allow us to better assess the strengths of different clustering paradigms in this fuzzy grouping context.

FOG represents a significant movement forward in our ability to identify groups as it enables the location of fuzzy groups. Fuzzy groups are a more natural and compelling way of thinking of human social groups. An unintended consequence of this approach is that the strength of membership in groups and the prevalence of exclusive members are diagnostic. We saw historical case study evidence that the strength of membership was valuable in predicting the willingness of actors to act with their group; e.g., in the case of the Sampson data, the strength of group membership is a good indicator of the order of leaving. We saw similar evidence that the higher the prevalence of exclusively tied individuals the higher the likelihood that the group would fission into an isolated component; e.g., in the case of the DGG group 2 is predominantly composed of exclusive members and it is the group that ultimately fissioned off. While preliminary, and based on only two case studies, these findings are strongly suggestive. As such, we expect that a fuzzy group approach may be key to building a mathematics of emergent group phenomena.

Acknowledgments

This work was supported in part by the National Science Foundation under the IGERT program (DGE-9972762) for training and research in CASOS, and the Office of Naval Research under Dynamic Network Analysis program (N00014-02-1-0973, ONRN00014-06-1-0921, ONRN00014-06-1-0104). Additional support was provided by CASOS – the Center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied of the National Science Foundation, the Office of Naval Research, or the U.S. government.

References

Airoldi, E., Blei, D., Xing, E., Fienberg, S., 2005. A latent mixed membership model for relational data. In: Proceedings of the ACM Link-KDD Workshop in conjunction with ACM SIG-KDD.

Airoldi, E., Blei, D., Xing, E., Fienberg, S., 2006. Mixed membership stochastic block models for relational data with application to protein-protein interactions. In: Proceedings of ENAR Annual Meetings.

Battacharya, I., Getoor, L., 2004. Deduplication and Group Detection Using Links. KDD Workshop on Link Analysis and Group Detection.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

Borgatti, S.P., Everett, M.G., Freeman, L.C., 2005. UCINET 6. Analytic Technologies.

Breiger, R., Boorman, S., Arabie, P., 1975. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. Journal of Mathematical Psychology 12, 328–383.

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, 1024–1031.

Clauset, A., Newman, M.E.J., Moore, C., 2004. Finding community structure in very large networks. Physical Review E, 70.

Davis, A., Gardner, B.B., Gardner, M.R., 1941. Deep South: A Sociological and Anthropological Study of Caste and Class.

Diesner, J., Frantz, T., Carley, K.M., 2006. Communication networks from the Enron email corpus. Journal of Computational and Mathematical Organization Theory 11, 201–228.

Doreian, P., Mrvar, A., 1996. A partitioning approach to structural balance. Social Networks 18, 149–168.

Doreian, P., Batagelj, J., Ferligoj, A., 2005. Generalized Blockmodeling. Cambridge University Press.

Freeman, L.C., 1992. The sociological concept of 'Group': an empirical test of two models. American Journal of Sociology 98, 55–79.

Girvan, M., Newman, M.E.J., 2002. Community structure in social and biological networks. In: Proceedings of the National Academy of Sciences USA, 99, pp. 7821–7826.

Kashima, H., Tsuboi, Y., 2004. Kernel-based discriminative algorithms for labeling sequences, trees, and graphs. In: Proceedings of 21st International Conference on Machine Learning.

Kubica, J., Moore, A., Cohn, D., Schneider, J., 2003a. cGraph: a fast graph-based method for link analysis and queries. In: Third Workshop on Link Analysis, Counterterrorism and Security, SIAM National Conference on Data Privacy.

Kubica, J., Moore, A., Schneider, J., 2003b. Tractable Group Detection on Large Link Data Sets. IJCAI Text-Mining and Link-Analysis Workshop.

Lorrain, F., White, H.C., 1971. Structural equivalence of individuals in social networks. Journal of Mathematical Sociology 1, 49–80.

Moody, J., White, D.R., 2003. Social cohesion and embeddedness: a hierarchical conception of social groups. American Sociological Review 8 (1), 1–25.

Newman, M.E.J., 2004a. Analysis of weighted networks. Physical Review E 70, 056131.

Newman, M.E.J., 2004b. Coauthorship networks and patterns of scientific collaboration. In: Proceedings of the National Academy of Sciences USA 101, pp. 5200–5205.

Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Physical Review E, 69.

Page, L., Brin, S., 1998. The Anatomy of a large-scale hypertextual (Web) search engine. Computer Networks and ISDN Systems 30, 107–117.

Palla, G., Derényi, I., Farkas, I., Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814–818.

Sampson, S.F., 1968. A novitiate in a period of change: an experimental and case study of social relationships. Unpublished Doctoral Dissertation.