Social Networks 55 (2018) 1–10

Contents lists available at ScienceDirect

Social Networks

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

Patterns of co-membership: Techniques for identifying subgraph composition

Sean M. Fitzhugh *,[1], Carter T. Butts [1,2]

3151 Social Science Plaza, University of California, Irvine, CA 92697, USA

ARTICLE INFO

Article history:

Keywords:
Cohesive subgroups
Focus theory
Exploratory data analysis
Homophily
Cliques
Friendship networks
Facebook

ABSTRACT

Different social processes give rise to network structures with distinctive properties. In this paper our goal is to identify the social processes that give rise to distinct network structures (specifically, subgroups). We examine particular structural meta-relations by identifying the properties of individuals associated with specific subgroups. Clues to the process of group formation and the context in which these groups form and persist may be extracted from the properties of individuals in those groups. Following this intuition, we propose a general technique for identifying systematic patterns of attribute occupancy to determine how individual attributes may drive group formation. To connect the social context in which groups form to their structural signatures, we relate subgroup composition to nodal attributes. We illustrate the utility of comparing subgroup (e.g., clique, n-clique, k-core, etc.) co-membership with nodal co-membership in a variety of attributes. The correlations between these two co-membership matrices illustrate clearly the strength of association between shared attributes and shared subgraph membership. Furthermore, examining these correlations across groups of different sizes indicates where these attributes are most strongly associated with group co-membership. Additionally, these correlations fit well into a QAP framework to determine where shared subgraph membership has a stronger (or weaker) relation to shared attribute membership than we would expect by chance. We demonstrate the technique with a series of large, online friendship networks on the order of thousands of nodes to illustrate how factors such as gender, cohort, residence, and other attributes are associated with co-membership across a range of clique sizes.

Published by Elsevier B.V.

1. Determinants of network structure

Observed networks represent the structural influence of countless social processes, including homophily (McPherson et al., 2001; Newman, 2002), propinquity (Bossard, 1932; Brakman et al., 1999; Festinger et al., 1950), preferential attachment (Price and de Sola, 1976; Barabási and Albert, 1999), disassortative mixing (Johnson et al., 2010; Newman, 2002, 2003), and many others. These processes rarely occur in isolation, as they typically act in tandem to shape the web of relations represented by the network. Disentangling this multitude of social processes in networks has been an ongoing topic of interest in the field and is a particular challenge for characterizing large graphs.

While networks may be driven by a large number of competing social processes acting at multiple scales (Contractor et al., 2006), many of the early structural theories focused largely on the determinants of lower-order properties of graphs and/or processes arising in small group settings. Among the earliest theories in social network analysis are balance theory (Cartwright and Harary, 1956; Heider, 1958; Newcomb, 1953; Davis, 1967), theories of triadic closure or transitivity (Holland and Leinhardt, 1971; Davis, 1967), and theories of reciprocity (Bronfenbrenner, 1943; Katz and Powell, 1955). These theories reflect where the field and its data originated. That is, they tend to focus on relatively simple, homogeneous rules that govern processes that were originally observed in relatively small networks in which all members could be aware of and interact with one another (i.e., fewer than one hundred nodes or so). Several of the earliest studies of networks came from studies of small groups in psychology (Moreno, 1934), communication (Bavelas, 1950; Newcomb, 1953; Leavitt, 1951), and anthropology (Barnes, 1954; Mitchell, 1974), as well as other relatively small,

* Corresponding author.
E-mail addresses: sean.fitzhugh@uci.edu (S.M. Fitzhugh), buttsc@uci.edu (C.T. Butts).
[1] Department of Sociology, University of California, Irvine, USA.
[2] Institute for Mathematical Behavioral Sciences, Department of Statistics, Department of Electrical Engineering and Computer Science, University of California, Irvine, USA.

https://doi.org/10.1016/j.socnet.2018.03.006
0378-8733/Published by Elsevier B.V.

homogeneous, well bounded groups (Coleman et al., 1957). Reflecting the field's origins, these theories explain well the processes driving tie formation in relatively homogeneous networks with few restrictions on mutual awareness and potential interaction. As networks grow in size from tens to hundreds to thousands of nodes and beyond, however, we struggle to glean insight into the network's structure with lower-order properties such as transitivity and reciprocity.

Measures based on those lower-order network properties have long served an important function in the analysis of networks. Triadic closure, degree distribution, and network centralization, for example, have long aided in exploratory, data analytic characterizations of networks. Such analyses are essential to obtain a general understanding of a graph's structure or properties, often in support of generating hypotheses about the drivers of the social phenomena underlying that graph's structure. These types of exploratory analyses play a role as vital precursors to classical hypothesis-testing methods such as brokerage scores (Gould and Fernandez, 1989), quadratic assignment procedure (Hubert and Arabie, 1989), conditional uniform graph tests (Anderson et al., 1999), and more recent, model-based hypothesis-testing methods such as exponential-family random graph models (Robins et al., 2007; Hunter et al., 2008), stochastic actor-oriented models (Snijders, 2001, 2017; Snijders et al., 2010), and relational event models (Butts, 2008). This paper follows in the tradition of using exploratory analysis to characterize network structure. However, we aim to develop an approach that provides insight into the structure of large graphs, where lower-order properties struggle to provide useful insight and where coping with heterogeneity becomes increasingly important.

With the increasing availability of large-scale network data (i.e. on the order of thousands, tens of thousands, hundreds of thousands of nodes, or beyond), the field has increasingly begun to grapple with the theoretical and methodological challenges of working with large-scale networks representing relations such as friendships on Facebook, citations among academic publications, co-authorship collaborations, emails exchanged within large companies, and physical proximity measured via sensors. Large graphs are more structurally complex and have more interdependent "moving parts," which complicates the process of identifying individual forces driving tie formation. Describing local structural configurations and testing traditional network theories have been ongoing challenges in these kinds of networks. For example, methodological innovations by Willer et al. (2012) were motivated by a lack of techniques that could test network exchange theories in networks larger than twelve. In another example, Leskovec et al. (2008) found that models based solely on theories of preferential attachment are inadequate to reproduce the community structure observed in large networks (on the order of tens of thousands to millions of nodes) such as online social networks, email communication networks, and citation networks. These examples serve as cautionary tales that large networks are particularly difficult to analyze because their structure may be driven by a variety of co-occurring processes. Harnessing classical theories to explain the structure of large-scale networks is an ongoing challenge in the field and one that forces us to determine how to disentangle the many processes that drive the formation of such networks. Describing local structure in large networks is another ongoing challenge, which some have addressed using network motifs, recurring subgraphs that typically range in size from two to five. Milo et al. (2002, 2004) have used these motifs to characterize local structure in worldwide web networks, language networks, networks of positive sentiment among prison inmates, and friendship relations among college freshmen. Because large graphs reflect a wide variety of social processes, the field has adapted its approach to using traditional theories and for identifying local structure. In this paper we continue following this approach of adapting how we utilize classical theories to explain network structure by using Feld's focus theory to motivate a family of techniques that identifies factors associated with the determinants of group co-membership within networks.

2. Identifying sites of group formation

We build on Scott Feld's *focus theory* in order to identify the factors driving subgroup formation. Feld (1981) argues that individuals organize their relations around *foci*, sources of joint activity such as voluntary organizations, workplaces, and neighborhoods. Individuals in these shared spaces are more likely to interact and subsequently form ties with each other. These foci serve as *nucleation sites*, spaces where ties develop and become reinforced. Feld's focus theory scales to large networks, as Feld describes how larger populations typically have larger numbers of foci. Because members of the population have bounded rationality (Slovic et al., 1974) and cannot put forth the effort to be simultaneously involved in all foci, these foci tend to occur at specific socio/demographic/spatial "locales" in the population (i.e., within Blau space – see McPherson, 1983), such as workplaces, neighborhoods, and voluntary associations. Individuals utilize a limited number of foci and these foci drive relationship formation in specific parts of the population. To identify the determinants of tie formation in subgraphs, we build on this concept of shared features as sites for tie formation.

Large networks typically represent a conglomeration of social processes, although these processes do not necessarily operate homogeneously throughout the network. Rather, different social processes often operate on different scales. Some nucleation sites for tie formation are constrained to very specific, local scales: for example, living in the same neighborhood. Other processes, such as tie formation driven by gender or racial homophily, permeate throughout society. Some nucleation sites give rise to large, dense groups while others create much smaller groups. For example, communication ties among individuals co-residing in a household form subgroups that look quite distinct from subgroups of communication ties of co-workers in a multifaceted organization with hundreds of employees. In this paper we look for signals of different social processes based on the scale of these subgroups. We do this by using shared attributes to examine the composition of subgroups across different sizes. Shared attributes among individuals in a subgraph may lend insight into the nucleation site or sites that have contributed to the formation of ties among them. This notion of different mechanisms or mixing patterns giving rise to distinct substructures with differential frequency goes back several decades in the social network literature (Frank and Strauss, 1986; Davis, 1979; Holland and Leinhardt, 1976), but scalable exploratory methods capable of linking such substructures with particular foci are still poorly developed. Identifying the characteristics of individuals composing these subgraphs will help to uncover the forces that shape these networks.

3. Linking group membership to shared attributes

To identify which foci are associated with group formation we introduce a general family of techniques for relating group co-membership to shared attributes. We define "group" here generically to be any class of structural subgroup such as cohesive subgroup (clique, n-clique, or k-core, etc.) as appropriate for the social process of interest. Having chosen a subgroup definition, we then identify co-membership in these subgroups. We track subgroup co-membership across all observed subgroup sizes, as the forces driving subgraph formation may vary across the range of subgroup sizes. For example, the factors driving co-membership in maximal cliques of size 4 may differ from factors driving co-