Finding Overlapping Communities in Social Networks

Mark Goldberg*, Stephen Kelley†, Malik Magdon-Ismail*, Konstantin Mertsalov *, and Al Wallace ‡

* *Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY*
{goldberg, magdon, mertsk2}@cs.rpi.edu
† *Cyberspace Sciences and Information Intelligence Research Group, Oak Ridge National Laboratory, Oak Ridge, TN*
kelleys@ornl.gov
‡ *Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY*
wallaw@rpi.edu

***Abstract*—Increasingly, methods to identify community structure in networks have been proposed which allow groups to overlap. These methods have taken a variety of forms, resulting in a lack of consensus as to what characteristics overlapping communities should have. Furthermore, overlapping community detection algorithms have been justified using intuitive arguments, rather than quantitative observations. This lack of consensus and empirical justification has limited the adoption of methods which identify overlapping communities. In this text, we distil from previous literature a minimal set of axioms which overlapping communities should satisfy. Additionally, we modify a previously published algorithm, Iterative Scan, to ensure that these properties are met. By analyzing the community structure of a large blog network, we present both structural and attribute based verification that overlapping communities naturally and frequently occur.**

***Keywords*-social network analysis, community detection, overlapping groups**

I. Introduction

The advent of the Information Age has opened new possibilities in the field of social network analysis by making very large repositories of data available to researchers. Phone calls, electronic communication via email, and scientific publication co-authorship records are now stored in centralized, relatively easily-accessible locations. In addition, many social networking services and blog providers have emerged as important forums for individual expression and discourse. All of these provide researchers with rich and publicly observable data to use in the analysis of social interactions.

As social networks grew to sizes far beyond the possibility of manual processing, it became increasingly important to develop computationally efficient and accurate algorithms that can bring important features of a network to the forefront. The identification of communities is essential to understanding of the structure and functionality of large networks. Previous work has used such analyses to identify voting blocs [18], protein families [9], and scientific communities [17].

In this paper, we address two issues facing the field of community detection: **(1)** the lack of a well accepted definition of what constitutes a community; and **(2)** a quantitative analysis of large-scale social networks demonstrating significant numbers of communities which have non-trivial overlap.

Existing literature on locating communities in networks contains various definitions of what constitutes a community. A thorough survey on the field can be found in [8]. The definitions range from maximal complete sub-graphs in the network to sets comprised of individuals who are more similar to each other than to outsiders. However, a typical way of defining communities is to design an algorithm that outputs sets and declare the output of the method to be communities. Thus, different algorithms, or even variations of algorithms, yield different sets of communities in a network. The lack of a unified definition of a community makes it difficult to fairly compare the performance of different community detection algorithms.

Much of the current work treats the problem of locating communities as a hierarchical partitioning problem ([5], [7], [9], [13], [14], [19], [20]). According to this approach, the community structure of a network is assumed to be hierarchical; individuals form disjoint groups which become subgroups of larger groups until one group, comprising the whole society, is formed. While this assumption is valid for some types of networks, *e.g.*, organizational networks or taxonomies, many social networks contain pairs of communities that overlap while not containing each other as a sub-community. Individuals often associate across many different social circles, such as those focused around the workplace, family unit, religious group, or social club. In this case, assuming the hierarchical social structure of the network would lead to missing important information about members' attachment to the numerous social circles with which they concurrently interact.

The observations above have been used as intuitive justification for designing algorithms that find overlapping communities in social networks. These algorithms have taken a number of approaches including methods which are based on discovering specific substructures in the network [17], identifying edges with high betweenness ([11], [12]),

maximizing modified versions of modularity [15], locating locally dense sets of vertices [1]–[4], and deriving fuzzy communities via probabilistic models which best explain observed edges in the network [6].

We formulate two simple properties, axioms, for a set of members to qualify as a community. These minimal requirements are often (but not always) satisfied by the definitions currently in use. We give only the minimal, in our view self-evident, requirements in order to preserve the flexibility and generality of the definition. In particular, we do not define communities as the output of a specific algorithm. Rather, we specify only a set of guidelines for what should constitute a community.

Our definition is local: a set of members may qualify as a community independent of both the total communication intensity in the network and their membership in other communities. This independence with respect to membership in other communities restricts the application of our axioms to those methods where an individual's membership in a group is binary. Methods which attempt to discover fuzzy community assignments violate this independence. Additionally, it is worth noting that the axioms do not explicitly refer to the issue of community overlap; this opens up the possibility for some communities to overlap, but does not require it. Computational experiments show that this possibility is often realized in social networks.

The minimality of the requirements may lead to implementation difficulties when the number of all sets satisfying the axioms is large. Because of this, depending on specifics of the application at hand, filtering out some candidate-sets based on auxiliary constraints might be needed.

In the second part of the paper, we empirically show, through quantitative experiments, that a definition of communities allowing for overlap is essential for analysis of social networks. We empirically analyze several social networks, including a small, commonly used benchmark dataset, Zachary's Karate Club [22], and a large, real-life dataset, the network of communication in the blog-provider LiveJournal [10]. We present a heuristic algorithm which outputs a collection of communities that satisfy our axioms. We further demonstrate that, in real-life social networks, a large number of individuals are members of communities which have a non-trivial overlap with other communities. Using structural properties of communities identified by our overlapping group detection algorithm and the declared friendship relations of the underlying network, we demonstrate that a significant number of the associations are not captured by any set of disjoint communities.

The graph theory definitions not introduced in this paper can be found in [21].

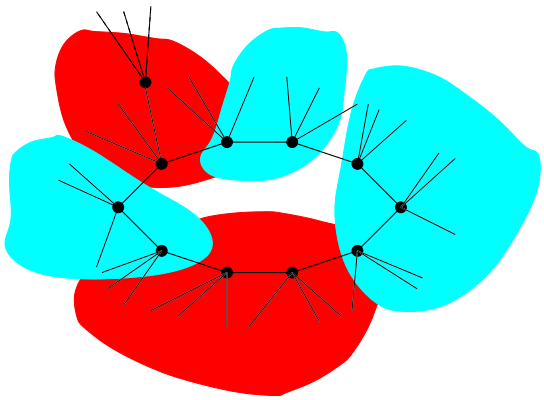

Figure 1. A demonstration of local optimality

II. Defining Communities

A social network is represented by a weighted graph $G = (V, W)$, where the edge weights w_{ij} measure[1] similarity of the nodes. For example, the edge weight between two books would be large if the same customer bought both books. An edge between two members of an on-line social network might exist if they communicated with each other. A "community" of books might represent a topic; a community of members on an on-line social network might represent a social group. Such communities are expected to overlap. Overlapping communities pose a problem for standard definitions of communities, as we will soon see. In the remainder of this paper, we will focus on communication networks (though our discussion is general), where communication can be viewed as a measure of similarity.

The starting point is a notion of a set-density, and for concreteness, we will use the density definition:

$$d(S) = \frac{W_{in}(S)}{W_{in}(S) + W_{out}(S)}, \tag{1}$$

where $W_{in}(S)$ is the total weight of edges whose endpoints are both in S, and $W_{out}(S)$ is the total weight of edges with one end point inside S and the other outside S. The rationale behind this notion of density is that it captures how much intra-group similarity there is compared with the similarity between S and the outside world.

It is typically understood that communities should display *more* intra-group similarity than extra-group similarity. This is self-evident for non-overlapping communities, but when communities are allowed to overlap, we have to re-examine even such basic intuitive proposition. To illustrate, consider the stylized example in figure 1. This figure depicts some form of organized/coordinated ring-group which would intuitively pass as a community (for example, a committee of NSF-reviewers). Since we allow overlapping groups, a

[1]For dynamic social networks one might consider a time-series of such graphs.

node could belong to multiple communities, as illustrated by the shaded areas. A node belongs simultaneously to this ring-community as well as to other communities. By virtue of belonging to those other communities, the node communicates extensively outside the ring-group (especially if the node belongs to many other communities). This means that the node displays *more extra-group* similarity than intra-group similarity. There is no flaw with the intuition that a community should display intra-group similarity; the reason the extra-group similarity can be larger is because the communities can overlap. Thus, we should rule out algorithms which search for communities for which $d(S)$ is larger than some threshold (for example, by requiring intra-group similarity to be larger than extra-group similarity implies that $d(S) > \frac{1}{2}$ for the specific density notion above). Note that the ring itself in our example, though it is *connected* and appears structured, is not particularly dense; in fact, if each member connects to δ external nodes, then $d(S) = 1/(\delta+1)$, which can be sufficiently small. Other communities may not have as low a density as this. We can go farther in claiming that this subset should be considered a community independent of the nature of the other communities in the network. Accepting the *locality* property of the communities suggests that the methods which define a global objective function (for example, modularity [14]) and optimize it to identify all the communities might fail to discover the ring-community. Such methods have found success in partitioning a network, but when overlap is allowed and essential, it is not even clear how to properly define global objective functions.

It is useful to consider one of the algorithms proposed in the literature for finding overlapping communities: the clique percolation method [17]. In a nutshell, the algorithm first finds all cliques of size k, and defines the k-clique graph whose nodes are the k-cliques, and two nodes are adjacent if the corresponding cliques share $k-1$ nodes. The connected components of the k-clique graph are used to define the communities in the network: the nodes in the union of the k-cliques which correspond to a connected component are declared to be a community. For $k = 2$, clique percolation defines the communities as the connected components in the network. It would be hard to argue that, for reasonably sized k, a community so defined would satisfy most intuitive expectations of a community; the problem with this definition is that it sets up a very rigid definition for a community, not much weaker than requiring the community to be a clique–if one edge is missing, or if two k-cliques overlap by only $k-2$ nodes, then it is not acceptable. Clique percolation would not, for example, be able to find the group illustrated in our toy problem above. The main problem with such a definition is that it is too rigid, and is uniform over the whole network, requiring all communities to "look the same."

As was already mentioned about our toy ring group, the density of our ring-community is $d(S) = 1/(\delta+1)$. One can easily verify that if we remove a node u from the group, its density drops to

$$d(S-u) = \frac{1}{\delta + 1 + \delta/(|S|-2)}.$$

Alternatively, suppose we try to add one of the neighboring nodes z to S. To illustrate, assume that this node has one connection into S and β connections to other nodes. In this case, adding z changes the density to

$$d(S+z) = \frac{1 + 1/|S|}{\delta + 1 + \beta/|S|},$$

which is smaller than $d(S)$, when z has more connections to the outside world than the average for nodes already in S. This means that S is *locally optimal* with respect to single node moves. Thus, the requirement of local optimality can capture S as a community. Further, many different types of communities can be locally optimal with varying densities, and locally optimal communities can overlap. Not being able to improve a community (as measured by the density d) is intuitive; this does *not* require a high density or a specific structure of the community. The unified idea of the discussion is that a community is a *locally* defined object. A community in one part of the network should not rely on what is going on in another part of the network. Further, community structure can vary over the network – communication in some communities can be more intense than in others; their structures can be different.

Community Axioms: We now state the minimum requirements of a community.

> **Connectedness.** A community should induce a connected sub-graph in the network. If the only path from one node to another in the community is via some external node, it suggests that the community is incomplete.
>
> **Local Optimality.** According to an appropriate density metric $d()$, predefined on all subsets of nodes, the density of a community cannot be improved with the removal or addition of a single node.

Note, that the local optimality requirement, but not the connectivity requirement, was first introduced in [2], [3]. Examples can be easily developed of locally optimal sets that induce disconnected sub-graphs. Our community axioms posit, in particular, that communities are identified "locally," within one-hop distance from the set. As we will see, these two axioms alone are sufficient for discovering communities which overlap, and satisfy the intuitive properties we expect of a community.

Algorithmically, it is not easy to identify all communities satisfying these properties, and so we resort to a simple heuristic which we discuss next. Our goal is to show that the communities discovered using this heuristic which satisfy the two community axioms reveal that overlap is essential

in social networks; this in turn means that one must use a definition of a community which allows for overlap and addresses all the issues discussed in this section.

III. Connected Iterative Scan

In [3], the authors describe a community detection algorithm, termed Iterative Scan. Here we describe a modification of Iterative Scan to discover communities satisfying both axioms.

Iterative Scan, which we will refer to as IS, consists of repeated "scans" each starting with an initial set developed by the previous scan (a "seed" set for the first iteration). It examines each node of the network, adding or removing it the density of the set is increased as a result. The scans are repeated until the set is locally optimal with respect to a defined density metric. The choice of the seed sets is not predetermined: they can be the nodes, or the edges of the network. A procedure for seeding, called Link Aggregate, is presented in [2]. Link Aggregate efficiently produces seed-sets that form a cover (with some overlaps) of the entire vertex set. The nodes are evaluated by IS in the order of increasing node-degree, from low to high degree. Iterative Scan in this form had been used for a variety of applications [10]. A similar method, implementing the idea of the greedy local optimization (as a replacement of a scan in IS) was later given in [1]. For every iteration, the algorithm examines all vertices in order to find the one which causes the maximum increase of the density. That vertex is used to update the current set.

The density metric itself can be defined in a number of ways; our analysis uses a modification of the standard density function in Equation 1. Rather than $\frac{W_{in}(S)}{W_{in}(S)+W_{out}(S)}$, we use $2W_{in}(S)$ for all instances of W_{in}. This density metric has been used an recent literature and has been shown to give good results [1]. Our experiments show that in many social networks, there is a very large set of potential communities, *i.e.*, sets that satisfy the two axioms above. Thus, the filtering out of candidate sets to be dictated by the specifics of the application domain is often necessary. One possibility is to order the candidates by $d(S)$, and consider as most "interesting" those communities which had more internal than external communication ($d(S) > \frac{1}{3}$). This filter is consistent with the notion of a "weak" community as defined by Raddicchi *et al* in [19] and is done to restrict the scope of the analysis for computational issues. Note that when overlap is allowed, this additional requirement might not be satisfied by all communities. The other possibility of filtering is to look at the communities for which $d(S) < \frac{1}{3}$, as these communities are still connected and locally optimal, even though their members communicate outside of the community a significant fraction of time, resulting in sparse internal communication.

To ensure the connectivity of the identified communities, we modify IS and termed the resulting algorithm Connected Iterative Scan, CIS. As is the case with IS, CIS consists of a number of scans that are repeated for a set until no change of the set occurs. Then the set is declared to be a community. Every scan proceeds through the nodes in the order of increasing node degree. Once a scan is finished, the set's connectivity is examined. If the set consists of multiple connected components, it is replaced by the connected component with the highest density, after which the next scan starts. Note that selecting only the highest density component effectively sidesteps the issue of repeatedly optimizing to the same, disconnected cluster. The seeding is this text is done using Link Aggregate.

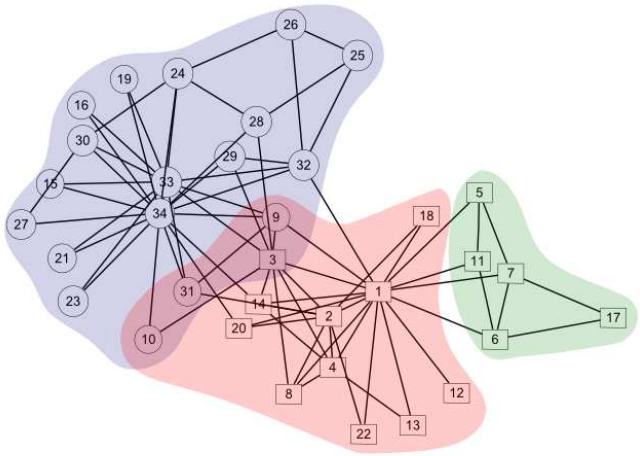

Figure 2. Overlapping groups found in Zachary's Karate Club data-set. Different shapes identify the eventual group division. Groups were ordered to correspond to the number of distinct seeds which produced them. Groups were then selected until the graph was covered. Additional examination of groups which are produced by fewer seeds offers insight into potentially overlapping subgroups of the primary groups presented here.

Sample results of CIS for a community analysis of Zachary's Karate Club data set [22] are given in Figure 2. In this case, the overlap of the two groups consists of the individuals who have equal number of associations with both communities. Additionally, the change in community size distribution due to the inclusion of the connectivity requirement is given in Figure 3 for the observed blog communication network, which will be introduced in the next section of this text.

The disadvantage of CIS is the same as that of IS; the methods may produce a large number of highly overlapping communities. However, this problem can be managed by effective post-processing of results and merging of highly similar communities. Looking at the Zachary karate club data, it is evident that the overlapping communities make intuitive sense.

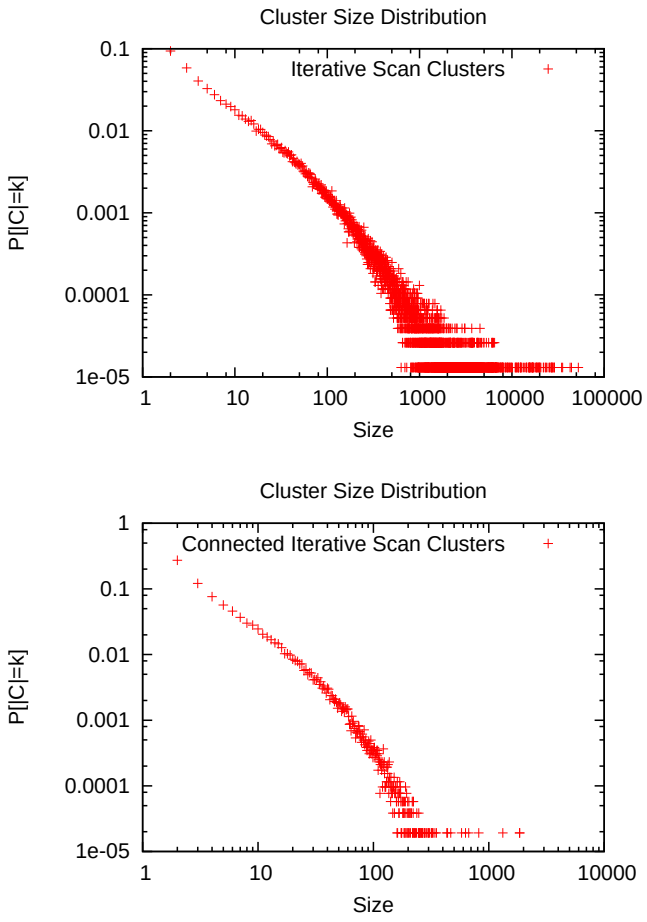

Figure 3. Plots demonstrating the size distribution for IS and CIS

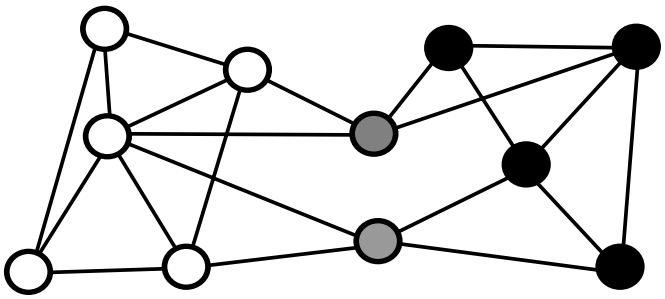

Figure 4. An example of a pair of groups that overlap. The overlap is identified by the gray vertices while individuals in only one group are colored black or white depending on the group of which they are a member.

IV. LiveJournal Data-set

We now consider a much larger social network, LiveJournal, on which we examine the existence of communities with overlap. It will thus be necessary to develop some quantitative methods for measuring the significance of overlap, since we will not be able to use visual validation.

In order to perform the analysis as described, the underlying network data needs to be composed of a communication network as well as user traits. LiveJournal provides services which allow for rich user-to-user interaction via blog postings, comments, friendships, and stated user interests. The data-set consists of user comment and interest records for the Russian section of this service over a 10-week period in 2008. An undirected network representing user comments is formed by placing an weighted edge between users A and B if A makes a comment in response to a post by B, with edge weight determined by the number of times user A comments on a unique post of B. This network is very large, consisting of over 300,000 users and 2.75 million weighted edges, with a total edge weight exceeding 5.6 million.

In addition to commenting on other users' posts, each individual in LiveJournal may declare which users he or she considers to be a "friend." This friendship relation is encouraged by the Friend-Feed feature, which presents new posts from any of a user's friends as soon as the user logs into the system. Since the relationship is not symmetric, it yields a directed graph with vertices whose in-degrees are, in general, different from their out-degrees. The distribution of the in-degrees, as well as out-degrees have presviously been shown to be scale free [10]. Small numbers of popular users collect comparatively large numbers of incoming friendship declarations while the vast majority of users collect little if any incoming friendship relations. These links will be used to determine the significance and similarity of groups and their overlaps. They will be explored further later in the text.

V. Significance of Overlap

In order to demonstrate that group overlap is a significant feature of some social networks, it is important first to consider the features which pairs of groups should have to indicate that the overlap between them is significant. Consider the overlapping groups presented in Figure 4. Here group A consists of white and gray vertices, and group B consists of the the black and gray vertices. By this definition, individuals represented by vertices colored gray are members of both group A and B.

For a pair of overlapping groups to have *significant* overlap, and thus be considered a *non-separable pair*, the groups and their overlap must fit certain criteria. In general, each criterion serves capture a different notion of quality which cannot be expressed via a single group (the union), or two, or three partitions. These criteria can be described conceptually as follows.

A. Structural Significance

The existence of overlap between a pair of groups should enhance the structural "quality" of each of the groups individually. For example, if the structural quality of each group is measured by one of the density functions introduced earlier, removing $A \cap B$ from A and B in the groups

expressed in Figure 4 would result in a decrease in the density of each group. The two vertices in the intersection $A \cap B$ have the same degree within each group as they have external to each group. Thus, relative to the previous density metric, the vertices should be a part of each group. Therefore, the overlap is the key to the structural significance of both groups in Figure 4.

B. Group Validity

It is also important that each group be somehow verifiable using a reasonable method relative to the input data. Ideally, using some underlying traits of the individuals in the network being analyzed, groups should have higher trait similarity between members than one would expect if membership in groups were determined at random. Examples of this type of validation have been used in various previous literature, using age and location as traits of the individuals [16]. Group validity is essential in filtering out groups that are products of random structures in the underlying communication graph, and serves to ensure that the group detection is accurate.

C. Overlap Validity

Using the same notion of trait similarity, the individuals within the overlap must have some similarity with the remainder of each group of which they are a member. In Figure 4, the graph is divided into three groups, $A - B$, $B - A$, and $A \cap B$ (white, black, and gray respectively). For overlap to be important, $A - B$ and $A \cap B$ must be similar, $B - A$ and $A \cap B$ must be similar, and $A - B$ and $B - A$ must be dissimilar, relative to certain significant traits in the data, that is individuals in the overlap need to be clearly similar to the remainder of either group. However, it is necessary that the remaining individuals in each group be dissimilar to those in the other group. If this dissimilarity does not exist, the overlapping pair can be captured in a single partition and overlap is not necessary to explain the relationships in the data.

Pairs of groups that satisfy each of these criteria are fundamentally sound communities due to their structural significance and their group validity. Conceptually, the existence of overlap validity restricts how the individuals can be placed in a partitioning. If all users of the three groups are placed in a single partition, dissimilar vertices in $A - B$ and $B - A$ are associated. If the vertices are placed in three partitions according to color, a strong association between $A \cap B$ and both $A - B$ and $B - A$ is missed. The vertices may be placed in a pair of disjoint groups only if the similarity between $A \cap B$ and both $A - B$ and $B - A$ is highly unbalanced. If the two similarities are comparable, however, one does not have justification to place the users in one group or the other. A detailed description of each of these cases is given further in the text. Significant numbers of non-separable pairs indicate that overlap is an essential component of communities within the network.

D. Measures

It becomes necessary to formulate a set of methodologies to indicate whether the notions of group validity and overlap validity are satisfied for a given community or pair of communities. We begin by identifying the set of data used in the analysis.

Due to the implementation of the Friend Feed provided by LiveJournal, friendship declarations can serve as an indicator of interest. By declaring a friendship, the declaring user is notified whenever his or her friend makes a post. It can be assumed that individuals which attract a large number of these friend declarations are highly important to the discourse on some set of topics. Thus, friendship declarations serve as a proxy for some set of declared interests from each user. In this analysis, an individual is defined as influential if he or she has a friendship in-degree of 300 or more. This criteria marks approximately 4,800 bloggers as influential.

The selection of a subset of the friendship relations was done for purely computational reasons, cutting the set of possible friend relations from 500,000 to 5,000. Additionally, interest declarations could be used as validation data. However, within LiveJournal, this data is entered via comma separated values, resulting in a much larger set of possible declarations. Additionally, the popular declared interests, such as ”books”, ”movies”, or ”music”, are much more universal than the most popular friendships. Further, words typed with spelling errors, abbreviations, slang, and the use of synonyms can all be indicative of the same set of topics. The friendship relationship is used in this situation because of its concreteness.

Now, given that each vertex i has a set of declared friendships F_i, we can describe our validation measures. The group validity requirement claims that there should be more similarity within the group than one would find at random. To measure this, we define the notion of *internal pairwise similarity* (denoted IPS). For a given community C, the internal pairwise similarity can be computed as

$$IPS(C) = \frac{\sum_{i \in C} \sum_{j \in C, j \neq i} J(F_i, F_j)}{|C|^2 - |C|}$$

where $J(F_i, F_j)$ is the Jaccard index between the two sets. Thus, the IPS value measures the average similarity between the friendship declarations of pairs within the community.

Revisiting the notion of overlap validity, it becomes apparent that a method comparing sets of friendship declarations are needed. Given a pair of overlapping communities A and B, three friendship declaration vectors can be computed. These vectors, denoted L_{A-B}, L_{B-A}, and $L_{A \cap B}$, give the probability that a vertex within each set indicated by the

subscript will declare a given individual in the popular friend set as a friend.

Once these vectors are constructed, the similarity between each of them can be calculated via the *cosine* similarity. Formally, this can be given, relative to two equal dimension vectors X and Y, as

$$\cos(\theta_{X,Y}) = \frac{X \cdot Y}{\|X\|\|Y\|} \tag{2}$$

A low value of $\cos(\theta_{X,Y})$ indicates that the vectors X and Y are close to orthogonal. High values indicate that the vectors have similar values across many dimensions.

Given the three friendship declaration vectors described previously, the *cosine* similarity between them can give an indication as to whether or not the overlapping group satisfies the overlap validity requirement. Namely, that the inter-group similarity $\cos(\theta_{L_{A-B},L_{B-A}})$ be less than the intra-group similarities $\cos(\theta_{L_{A-B},L_{A\cap B}})$ and $\cos(\theta_{L_{B-A},L_{A\cap B}})$.

In order to simplify this notion, the intra-group and inter-group similarities can be combined into a single statistic representing the relative similarity between the three sets. For the sake of notation, let the inter-group similarity $\cos(\theta_{L_{A-B},L_{B-A}})$ be given by the variable $inter$ and let each of the intra-group similarities $\cos(\theta_{LA-B,A\cap B})$ and $\cos(\theta_{L_{B-A,A\cap B}})$ be given by $intra_A$ and $intra_B$ respectively. These values can be combined into a measure of overlap validity as

$$\text{OV}(A,B) = \frac{intra_A + intra_B}{2} - inter$$

for values of OV(A, B)>0, the intersection is more similar to each group than the remainder of each group is with each other, indicating that the overlap is split in its association with each set.

VI. Results on LiveJournal

We applied the Connected Iterative Scan algorithm, CIS, to the LiveJournal data-set to produce a set of communities which satisfy the axioms. We also partitioned this graph using the algorithm CNM designed by Clauset, Newman, and Moore [5] to give the reader a point of reference and to demonstrate the difference in community sets produced by the two methods. Statistics detailing the number of groups, average size, average density, modularity (Q, only applicable for the partitioning), and the number of vertices which are placed in at least one community ("Cov") are given in Table I.

Table I
Statistics of groups discoved via CNM and CIS

	Groups	AvSize	AvDens	Q	Cov
CNM	264	1190	0.745	0.485	100%
CIS	14903	168.8	0.455	–	47.5%

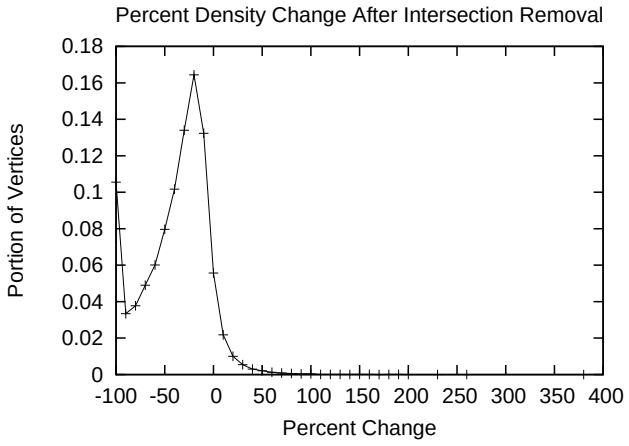

Figure 5. Portion of clusters that experience a given percentage change in density when the intersection of an overlapping pair is removed. Portions are collected in bins of size 10%. This plot contains 50 data points.

The partitioning produces a small number of sets across a wide variety of sizes, while the overlapping group detection produces a much larger number of smaller groups which do not cover the entire graph. Coverage is not a requirement; it is not necessary for every node to belong to a cluster. Rather, we are interested in finding those groups which naturally overlap and studying the significance of this overlap.

If the overlapping groups detected fit the requirement of having structural significance, removal of a pair's overlap will produce a decrease in group quality, as measured by the density d. Overlapping groups are more compelling when the overlap is structurally necessary for each group. After filtering out subset inclusion (a trivial form of overlap), the remaining overlapping groups display a high degree of structural significance for the overlap. Specifically, for 80.8% of the overlapping pairs, both groups in the pair experience a decrease in density if the intersection is removed. Figure 5 shows the distribution of changes in density when the overlap is removed. Even though we observed that some groups are improved by the removal of intersection, the overwhelming majority of groups experience a significant decrease in density. We conclude that, in general, community overlap is structurally significant.

We now investigate the validity of the groups found with respect to user traits. Figure 6(a) shows the average internal pairwise similarity between users within a community as well as the average similarity between users in connected random groups as a function of size. The figure shows that groups produced by CIS have much larger amounts of similarity between users than the random case for sizes greater than 10. This value appears lower than random for sizes less than 10 due to the number of groups which have undefined friendship declarations. The portion of these groups discovered by CIS and at random are given in Figure

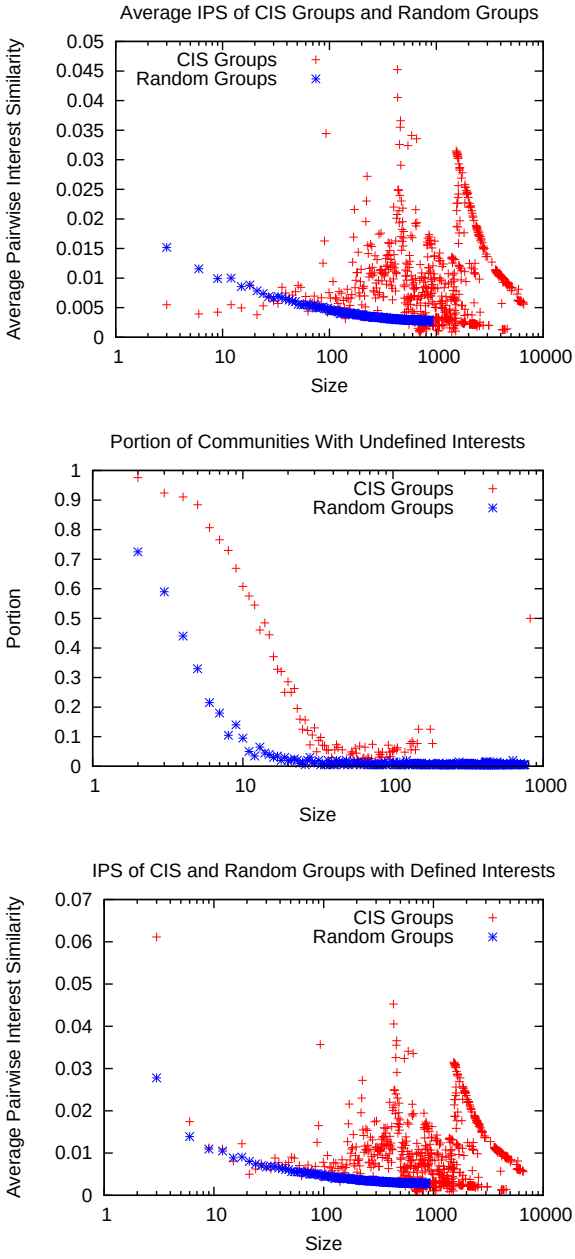

Figure 6. Plot showing the average pairwise Jaccard Index of vertex friendships for all pairs within discovered communities of the same size and values found in randomly generated connected groups of the same size. The plot indicates that there is more similarity in a majority of the discovered groups than one would expect at random.

6(b). Figure 6(c) shows the same information as Figure 6(a) but with these undefined friendships removed.

Figure 7 shows the overlap validity measure over pairs of groups with a given overlap. This value is compared with the overlap validity measure for randomly selected groups with the same size and overlap. The x-axis denotes the overlap of the pair, where overlap is defined as the Jaccard index of

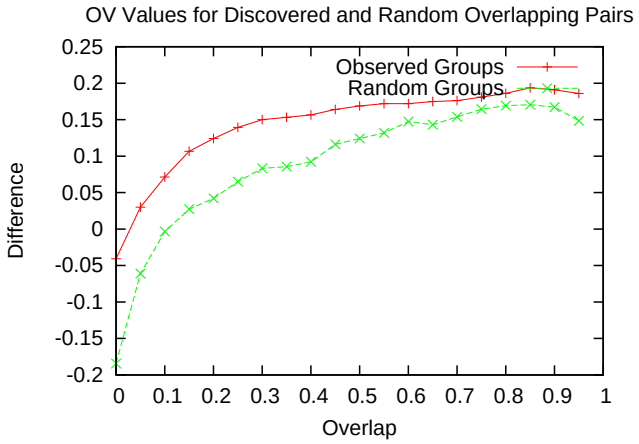

Figure 7. Curves showing the average overlap validity measure $\mathrm{OV}(A, B)$ for identified, non-subset overlapping pairs and random groups of the same size and overlap.

the two sets. Clearly, there is a larger difference in similarity between the groups identified via CIS and those generated at random.

For the 14,903 unique groups that were discovered, 6,373(30%) of them overlap with at least one other group such that the pair can be considered justified by the three conditions previously described. These pairs are composed of 125740 unique users, a very significant portion of the graph.

Further, a significant portion of the non-separable groups have comparable intra-group similarity between the intersection $A \cap B$ and both of the sets $B - A$ and $A - B$. If the similarities are considered comparable when they are within 5% of each other, 3,544 of the non-separable pairs have an overlap that is associated equally with the remainder of each group. These groups consist of 100,000 unique users. The existence of these groups is particularly significant in justifying overlap between communities. They clearly show that many sets of users are equally associated with distinct groups. Using a partition-based method for the detection of communities would either merge the entire pair into one group, failing to recognize the dissimilarity between the vertices in sets $A - B$ and $B - A$, or place the intersection with $A - B$ or $B - A$, missing the connection between the intersection and the other set.

VII. Conclusion

Previous attempts at developing algorithms for the detection of overlapping communities have been primarily intuitive, and were developed without first examining to what degree overlap occurs in naturally occurring networks. A large amount of justified overlap indicates that the added complexity of new methods is essential to capturing all relationships expressed in the data. As a test network, we

examined a social network composed of communications in a popular blogging service.

The overlap between groups must satisfy certain criteria to be considered significant. First, the inclusion of the common region in either group should enhance the quality of the groups by some metric. In addition, the groups themselves should be verifiable as significant through the use of a set of relevant user traits. Finally, the similarity between components of both groups involved in the overlap must be such that the intersection is more similar with the remainder of each group than the remainder of the groups are with each other. If each of these criteria is satisfied, placing the members of the group in some partitioning will not capture the subtle associations present in the data.

Building on the intuitive justification previously discussed in literature, this paper presents empirical evidence of the existence of a large amount of significant overlap in a network of blogs. The prevalence of such overlap demonstrates a deficiency in the traditional partitioning approach to discovering social communities. In order to capture such subtle associations, the notion overlap must be allowed and algorithms should take this into account.

VIII. Acknowledgments

This material is based upon work partially supported by the U.S. National Science Foundation (NSF) under Grant Nos. IIS-0621303, IIS-0522672, IIS-0324947, CNS-0323324, NSF IIS-0634875 and by the U.S. Office of Naval Research (ONR) Contract N00014-06-1-0466 and by the U.S. Department of Homeland Security (DHS) through the Center for Dynamic Data Analysis for Homeland Security administered through ONR grant number N00014-07-1-0150 to Rutgers University. This research is continuing through participation in the Network Science Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory under Agreement Number W911NF-09-2-0053.

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

The content of this paper does not necessarily reflect the position or policy of the U.S. Government, no official endorsement should be inferred or implied.

References

[1] J. K. A. Lancichinetti, S. Fortunato. Detecting the overlapping and hierarchical community structure of complex networks. *New Journal of Physics*, 11, 2009.

[2] J. Baumes, M. Goldberg, and M. Magdon-ismail. Efficient identification of overlapping communities. In *In IEEE International Conference on Intelligence and Security Informatics (ISI*, pages 27–36, 2005.

[3] J. Baumes, M. K. Goldberg, M. S. Krishnamoorthy, M. Magdon-Ismail, and N. Preston. Finding communities by clustering a graph into overlapping subgraphs. In N. Guimarães and P. T. Isaías, editors, *IADIS AC*, pages 97–104. IADIS, 2005.

[4] A. Clauset. Finding local community structure in networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 72(2):026132, 2005.

[5] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.

[6] G. B. Davis and K. M. Carley. Clearing the fog: Fuzzy, overlapping groups for social networks. *Social Networks*, 30(3):201 – 212, 2008.

[7] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72:027104, 2005.

[8] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.

[9] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc Natl Acad Sci USA*, 99(12):7821–6, 2002.

[10] M. Goldberg, S. Kelley, M. Magdon-Ismail, K. Mertsalov, and W. A. Wallace. Communication dynamics of blog networks. In *The 2nd SNA-KDD Workshop '08 (SNA-KDD'08)*, August 2008.

[11] S. Gregory. An algorithm to find overlapping community structure in networks. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007)*, pages 91–102. Springer-Verlag, September 2007.

[12] S. Gregory. A fast algorithm to find overlapping communities in networks. In *ECML PKDD '08: Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I*, pages 408–423, Berlin, Heidelberg, 2008. Springer-Verlag.

[13] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral. Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, 70(2):025101, 2004.

[14] M. E. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103(23):8577–8582, 2006.

[15] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri. Extending the definition of modularity to directed graphs with overlapping communities. *J.STAT.MECH.*, page P03024, 2009.

[16] G. Palla, A.-L. Barabasi, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

[17] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814, 2005.

[18] M. A. Porter, P. J. Mucha, M. Newman, and A. Friend. Community structure in the united states house of representatives. *Physica A: Statistical Mechanics and its Applications*, 386(1):414 – 438, 2007.

[19] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663, 2004.

[20] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[21] D. B. West. Introduction to graph theory. *Prentice Hall, Upper Saddle River, NJ*, 2003.

[22] W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.