# **Overlapping Communities in Social Networks**

Mark K. Goldberg CS Dept, Rensselaer 110 8th St. Troy, NY, USA goldberg@cs.rpi.edu Stephen Kelley Oak Ridge National Laboratory Oak Ridge, TN, USA kelleys@ornl.gov

Malik Magdon-Ismail CS Dept, Rensselaer 110 8th St. Troy, NY, USA magdon@cs.rpi.edu

Konstantin Mertsalov CS Dept, Rensselaer 110 8th St. Troy, NY, USA kmertsalov@gmail.com William A. Wallace CS Dept, Rensselaer 110 8th St. Troy, NY, USA wallaw@rpi.edu

# ABSTRACT

Identifying communities is essential for understanding the dynamics of a social network. The prevailing approach to the problem of community discovery is to partition the network into disjoint groups of members that exhibit a high degree of internal communication. This approach ignores the possibility that an individual may belong to two or more groups. Increasingly, researchers have begun to explore new methods which allow groups to overlap. One problem with existing approaches is that the definition of a community comes as the result of a particular algorithm. Such an approach to "defining" communities has been extended to overlapping communities with some success.

Our goals in this paper are twofold: first, to present an axiomatic approach to defining overlapping communities in terms of the properties a group should satisfy to be a community; and second, to justify the existence of overlapping in the structure of social communities experimentally using LiveJournal Blog data. Historically, the justification for overlapping groups has been primarily intuitive rather than quantitative. We present a heuristic algorithm which outputs a collection of communities that satisfy the required minimal properties and demonstrate that, in real-life social networks, a large number of individuals are members of communities which have non-trivial overlap with other communities.

## Keywords

social network analysis, community detection, overlapping groups

## 1. 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 a 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, accurate algorithms that can bring important features of a network to the forefront. Essential to the understanding of these networks is the identification of groups, or communities. Accurate group detection offers insight into the structure and function of large complex networks.

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

The existing literature on locating communities in a variety of different domains has led to various definitions of what constitutes a community, or a group. The definitions range from maximal complete subgraphs to sets comprised of individuals who are more similar to each other than to outsiders. In general, the definition of a community has been based on the output of some algorithm. The lack of a well accepted definition of a community makes it difficult to fairly compare the performance of different community detection algorithms, especially since these works define a community as the output of their algorithm

Much of the current work treats the problem of locating groups as a hierarchical partitioning problem (see [20, 10, 13, 6, 8, 12, 19, 9, 4]). 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. Consider an individual in a social network representing "friendship." He or she may have friendship relations 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 are being used as an intuitive justification for designing algorithms that find overlapping communities in social networks (see [3, 2, 5, 1, 16, 7, 14]).

We first consider the definition of overlapping communities. We formulate minimal properties (axioms) for a set of members to qualify as a community. These are minimal requirements that often appear in the definitions in current use. We attempt to give only the minimal requirements which preserves flexibility and generality. The starting point is a density measure defined on subsets of the vertices. Typically, the density function would represent the communication intensity in the network. The minimality of the requirements outlined by the axioms may lead to implementation difficulties when the number of all sets satisfying the axioms is too big. Because of this possibility, we acknowledge that depending on the specific application, filtering out of some candidate sets based on auxiliary constraints might be needed.

We then show empirically 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 ([21]), and a large, real-life dataset, the network of communication in the blog-provider LiveJournal ([11]). 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 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 if one restricts to disjoint communities.

#### 2. DEFINING COMMUNITIES

A social network is a weighted graph G = (V, W), where the edge weights  $w_{ij}$  measure similarity. <sup>1</sup> 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 online social network might exist if they communicated with each other. A "community" of books might represent a topic; a community of members on an online 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 typically some 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 a self-evident intuition for non-overlapping communities, but when communities are allowed to overlap, we have to reexamine even such a basic intuition. To illustrate, consider the stylized example below.



The idea in this picture is to depict 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 node could belong to multiple communities, as illustrated by the shaded circles. A node belongs simultaneously to this ring-community as well as to other communities. By virtue of belonging to those other communities, the node will communicate extensively outside the ring-group (especially if the node belongs to many other communities). This means that the node will display more extra-group similarity than intra-group similarity. There is no flaw with the intuition that a community should display intra-group similarity; it is because communities can overlap that the extra-group similarity can be larger. Thus, we can rule out algorithms which search for communities for which d(S) is larger than some threshold (for example requiring intra-group similarity to be larger than extra-group similarity means that  $d(S) > \frac{1}{2}$ . Note that the ring in our example, though it is *connected* and appears structured, is not particularly dense; in fact, if each member connects to  $\delta$  external nodes, then  $d(S) = 1/(\delta+1)$ , which can be arbitrarily small. Other communities may not have as low a density as this. We can go farther in saying that this subset should be considered a community independent of the natures of the other communities in the network.

<sup>&</sup>lt;sup>1</sup>If a social network is dynamic then one might have a time series of such graphs; we consider here a static snapshot of a network.

Thus, a community is a *locally* defined object. Thus, methods which define a global objective function (for example modularity [13, 4]) which is optimized to identify all the communities would fail this locality property. Such methods have found success in partitioning a network, but when overlap is allowed, it is not even clear how to define such a global objective function.

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 vertices are the k-cliques, and two vertices are are adjacent if the corresponding cliques share k-1 vertices. The connected components of the k-clique graph define the communities in the network; the nodes in the union of the k-cliques which correspond to a connected component are the 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 milder 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 already mentioned about our toy ring group, the density 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. For illustration, assume that this node has one connection into S and  $\beta$  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 community 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; etc.

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

- **Connectedness.** A community should induce a connected subgraph in the network if the only way to get 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.<sup>2</sup>

Our community axioms posit, in particular, that communities are identified "locally," within one-hop distance from the set. These two requirements are the only requirements we will impose. As we will see, these requirements alone are sufficient for discovering communities which overlap, and satisfy the intuitive properties we expect of a commuity.

Algorithmically, it is not easy to identify all communities satisfying these properties, and so we resort to a simple greedy heuristic, which we discuss next. Our goal is to show that the communities discovered using this greedy heuristic which satisfy the two community axioms already reveal that overlap is essential in social networks; this in turn means that one must use a definition of a community which allows overlap and addresses all the issues discussed in this section.

## 2.1 Connected Iterative Scan

To demonstrate the effectiveness of these axioms at discovering *overlapping* communities, any algorithm would do, so long as the groups produced are fit the axioms formulated above. We will use a modification of an algorithm presented in [3] to discover communities satisfying our two axioms.

In [3], the authors present group detection algorithm Iterative Scan. This algorithm, starting from a seed community, will add or remove vertices iteratively until the group is locally optimal with respect to a defined density metric. Vertices are evaluated in order of increasing degree, reconsidering vertices from low degree to high repetitively. This algorithm has been previously used for a variety of applications with interesting results (see [11]). A similar method based on greedy local optimization was also given in [1].

The density metric itself can be defined any number of ways, however, our analysis uses the standard density function in Equation 1. Our experiments show that in many social networks, there are 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 might be necessary. We chose to order them by d(S), and considered as most "interesting" those communities which had more internal than external communication  $(d(S) > \frac{1}{2})$ . This filter is consistent with the notion of a "weak" community as defined by Raddicchi

<sup>&</sup>lt;sup>2</sup>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 subgraphs.



Figure 1: Overlapping groups found in Zachary's Karate Club dataset. 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.

in [19]. Note that this additional requirement should not be expected of all communities when overlap is allowed. Indeed, it would be interesting to look at the communities for which  $d(S) < \frac{1}{2}$ , as these communities are still locally optimal (cannot be improved by single node moves), nevertheless the spend a significant fraction of their communication energy ourside the group. These communities are of a different type, in that they are involved in overlap with many other communities and/or they are quite sparse.

To ensure the connectivity of the identified groups, we propose a new variation of this algorithm called Connected Iterative Scan. Given a set of users as an initial seed, the algorithm proceeds through users in order of increasing vertex degree. Each user is considered for addition to or removal from the set as appropriate with changes being made as the individuals are evaluated if the modification improves the group's density. Once every user has been considered, the set's connectivity is examined. If the set consists of multiple connected components, the set is replaced by the connected component with the highest density, after which the optimization restarts. Selecting only the highest density component effectively sidesteps the issue of repetitively optimizing to the same, disconnected cluster. The optimization is complete for a seed when no single addition or removal increases the density of the connected set.

For this application, seeding is done via LinkAggregate as presented in [2]. The algorithm efficiently produces seeds that form a cover of the entire vertex set. Sample results of this algorithm for a community analysis of Zachary's Karate Club data set [21] are given in Figure 1. In this case, the two groups overlap because some individuals have equal number of associations with both communities. The advantage of using this group definition is that it doesn't state any requirements for specific structural properties of groups. Compared to the Clique Percolation Method found in [16], which only finds groups that are composed of an set of overlapping k-cliques, this criteria will find groups with a variety of centralized and decentralized structures including cliques, stars, and chains. The disadvantage lies in the number of sets that can be considered groups. However, this can be managed by effective post-processing of results. For instance, in the analysis of Zachary's Karate Club dataset given above, locally optimal groups are ranked in order of the number of distinct seeds that produce them. Groups which are discovered more often from distinct seeds may indicate greater stability and with respect to this analysis, are given a higher rank. These groups are then selected in order of decreasing ranking until the entire graph is covered.

Looking at the Zachary karate club data, it is evident that the overlapping communities make sense. We now consider a much larger social network, LiveJournal, on which to validate the need for overlap. It will thus be necessary to develop some quantitavtive methods for measuring the significance of overlap, since we will not be able to use visual validation.

## 2.2 LiveJournal Dataset

In order to complete this analysis as described, the underlying network data needs to be composed of a communication network as well as user traits. LiveJournal provides a set of 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 of 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. The directed nature of this relationship as well as the Friend Feed feature results in a scale free distribution of friendship in-degree. Small numbers of popular users collect comparatively large numbers of incoming friendship declarations while the vast majority of users collect little to no 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.

# 3. 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 2. For the sake of notation, group A consists of white and grey vertices, and group B is composed of the the black and grey vertices. By



Figure 2: An example pair of groups that overlap. The overlap is identified by the grey vertices while individuals in only one group are colored black or white depending on the group of which they are a member.

this definition, individuals represented by vertices colored grey 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 a general sense, each criterion serves to identify quality overlapping groups that cannot be expressed via a single group (the union), a two, or a three partition. These criteria can be described conceptually as:

#### 3.1 Structural Significance

The existence of overlap between a pair of groups should enhance the "quality" of each of the groups individually. For example, if the quality of each group is measured by the ratio of edges internal to the group to those which are cut by the boundary of the group, removing  $A \cap B$  from A and B in the groups expressed in Figure 2 would result in a decrease in the quality 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 quality metric, the vertices should be a part of each group since they increase the numerator while holding the denominator constant. Therefore, the overlap is key to the structural significance of both of the groups in Figure 2.

#### 3.2 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 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 [15]. Group validity is essential in filtering out groups that are products of random structures in the underlying communication graph and serve to ensure that the group detection is accurate.

#### 3.3 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 2, the graph is divided into three groups A - B, B - A, and  $A \cap B$  (white, black, and grey 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 - Amust be dissimilar relative to certain significant traits in the data. That is to say, 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 - Aare 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 into 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.

# 3.4 Cohesiveness and Similarity

In sociological literature, group cohesion is defined as the driving force that brings a set of individuals together to form a group ([18]). Cohesive forces can take a number of different forms: individuals coming together to perform a task, the threat of external competition for resources, or the underlying similarity of members of the groups. In the previously described LiveJournal dataset, friendship declarations can be used to approximate this cohesion. Due to the Friend Feed provided by LiveJournal, friendship declarations are a clear 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.

To measure the cohesiveness of a group, we propose the following method. For the global society, compute a vector G where, for every influential blogger i,  $G_i$  is the observed probability that a randomly selected person will list i as a friend. For each group l, we can similarly define a local friendship vector  $L^l$  where for each influential blogger i,  $L_i^l$  is the probability that a randomly selected person from group l will list i as a friend. With these two vectors, a notion of cohesiveness can formally be defined as the cosine similarity between them.

similarity
$$(G, L^l) = \cos(\theta_{G, L^l}) = \frac{G \cdot L^l}{\|G\| \|L^l\|}$$
 (2)

A low value of similarity  $(G, L^{l})$  for a given group l implies that the local and global friendship vectors are orthogonal, indicating that the group has a collection of interests that appear at probabilities that are significantly different from those of the global population. We examined this measure relative to a cohesion of each member of a set of randomly selected groups of the same size. A large difference between the cohesion of an observed group and the expected cohesion of a random group of the same size indicates that the observed group has a cohesive set of interests different from the global population. This demonstrates that the group in question has a unifying set of traits among its members and serves as group validation. As additional validation, the average similarity between the friendship vectors of pairs of individuals within a given group can be compared to the expected similarity of random pairs selected from the entire graph. Communities with higher similarity between pairs than pairs selected from the whole graph have more cohesiveness between users.

Further, the concept of cohesion can be used to compare pairs of communities that overlap. Consider two overlapping groups A and B similar to the groups shown in Figure 2. The intersection of the groups is given by  $A \cap B$ . The set of vertices which exist only in A is given by A - B and the set of vertices which exist only in B is given by B - A. A group with justified overlap will have a stronger similarity between the intersection  $A \cap B$  and each set B - A and A - B than the similarity of the sets B - A and A - B. As previously defined, pairs of groups which match this criteria are called *non-separable pairs*. Significant numbers of non-separable pairs indicate the existence of important associations which are missed when a group analysis is restricted to a collection of disjoint sets. This can be quantified in an *overlap validity* measure as

$$OV(A, B) = \frac{1}{2} (sim(A \cap B, A - B) + sim(A \cap B, B - A))$$
$$-sim(A - B, B - A)$$

where both  $sim(A \cap B, A - B)$  and  $sim(A \cap B, B - A)$  are greater than sim(A - B, B - A)

#### 4. **RESULTS ON LIVEJOURNAL**

For the LiveJournal dataset we applied our Connected Iterative Scan (CIS) algorithm to produce a set of communities which satisfy the axioms, and as a point of reference, we also partitioned this graph using the algorithm (CNM) designed by Clauset, Newman, and Moore ([6]). The results of the a quantitative analysis of the significance of community overlap in the network are presented in Table 1. It shows the number, size, and coverage differences of the groups identified by CNM and CIS, respectively.

The partitioning (CNM) produces a small number of nonoverlapping groups across a wide variety of sizes, while CIS produces a much larger number of smaller (overlapping) groups which do not cover the entire graph. Note that cov-





Figure 3: 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.

Statistics of Groups Found 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%

```
Table 1: Statistics of groups from CNM and CIS
```

erage 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 groups which are a subset of a larger group (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 3 shows more details of the exact 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 the overlap is structurally significant.

We now investigate the validity of the groups found, with respect to user traits. Figure 4 shows the average pairwise similarity between users within a community as well as the average similarity between users in random groups, where similarity is defined as the Jaccard index between the two individual's friendship declarations. The figure shows that groups produced by CIS have much larger amounts of similarity between users than the random case. Further, group validity can be observed using the notion of cohe-





Figure 4: Plot showing the average Jaccard Index of vertex friendships for all pairs within discovered communities of the same size and values found in randomly generated 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. The plot contains 1279 points representing the average pariwise similarity between groups of a given size while the random data consists of 554 points showing the same value. The error bars in the random data represent  $\pm$  one standard deviation.

siveness presented earlier. The distribution of average cohesiveness for various sizes may be examined by using the set of groups generated via Connected Iterative Scan, groups resulting from a partitioning using CNM, and randomly selected groups. These values are plotted with respect to size in Figures 5 and 6.

From these plots, a number of conclusions can be drawn. First, both group detection algorithms produce communities that are more cohesive than one would expect for a random group of a given size. CNM produces groups with high levels of cohesion for sizes larger than 5000 users, while CIS produces many groups in the 100 to 5000 user size which are not found via CNM that have better than random cohesion. Both methods appear to find similar groups with high levels of cohesion in the 10-100 user size. The two previous plots show that the groups discovered are reasonable and, for a majority of the overlapping groups, overlap is an important structural component. The third significant feature is overlap validity. 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 the two sets. Clearly, there is a larger difference in similarity between the groups identified via CIS and those generated at random.

For the 14903 unique groups that were discovered,  $6373(\sim 30\%)$  of them overlap with at least one other group such that the pair can be considered non-separable by the three validity conditions described previously. These pairs are com-



Figure 5: A plot demonstrating the cosine similarity between the interest vector for groups identified via CNM and CIS and the global interest vector. Each red point shows the size and similarity of a group discovered via CIS, while blue points represent groups found via CNM. The green lines represent the average cosine similarity of a randomly selected group of the same size plus and minus one standard deviation. Smaller values along the y-axis indiate significant differences from the global interest distribution. A simplified version of this plot showing the average cosine similarity across all sizes is given in figure 6.

posed of 125740 unique users, a very significant portion of the graph.

Further, a significant portion of the non-separable groups have comparable 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, 3544 of the non-separable pairs have an overlap that is associated equally with the remainder of each group. These groups consist of  $\sim 100,000$  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 placing the intersection with A-B or B-A, missing the connection between the intersection and the other set.

#### 5. CONCLUSION

Previous attempts at developing algorithms for the detection of overlapping communities have been primarily intuitive. These methods have been developed without first examining to what degree overlap occurs naturally in networks. A large amount of non-separable overlap would indicate that the added complexity of new methods which allow for overlap is essential to capturing all relationships expressed in the data. As a test network, we have examined a social network composed of communication on a popular blogging service LiveJournal. We have shown empirically that there



Figure 6: A plot demonstrating the average consine similarity between the interest vector of groups identified via CNM and CIS and the global interest vector. Each red point shows the average similarity of all groups of that size found via CIS, while the blue points represent the same values for groups discovered via CNM. The green lines represent the average cosine similarity of a randomly selected group of the same size plus and minus one standard deviation. Smaller values along the y-axis indicate significant differences from the global interest distribution. A plot showing all points is given in Figure 5

are many groups, composed of many users which overlap in a non-separable way – removing the overlap would decrease the quality of the groupings substantially. Disregarding this overlap would throw away the subtle relationships between different social communities, a fundamental aspect to the functioning of social networks.

In performing this empirical study, we developed methods for identifying significant (non-separable) overlap. For the overlap between groups to be considered significant, it must satisfy certain criteria. First, the inclusion of the common region into 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.

We showed that two commuity axioms, connectivity and local optimality, are enough to extract groups with significant overlap. In fact the definition is so flexible that it can find groups of very different forms. In building algorithms for social networks, it is essential to allow for overlap and for groups to be locally and non-uniformly defined. Our community axioms achieve this, and a simple greedy heuristic to find sets which satisfy these axioms appears to work well in practice.

Difference between Intragroup Similarity and Intergroup Similarity



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

# 6. 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.

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.

#### 7. REFERENCES

- 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] J. W. Berry, B. Hendrickson, R. A. LaViolette, and C. A. Phillips. Tolerating the community detection resolution limit with edge weighting. http://arxiv.org/PS\_cache/arxiv/pdf/0903/0903.1072v2.pdf, 2010.
- [5] A. Clauset. Finding local community structure in



Figure 8: The distribution of group sizes produced by Connected Iterative Scan in the LiveJournal dataset.

networks. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), 72(2):026132, 2005.

- [6] A. Clauset, C. Moore, and M. E. J. Newman. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.
- [7] G. B. Davis and K. M. Carley. Clearing the fog: Fuzzy, overlapping groups for social networks. *Social Networks*, 30(3):201 – 212, 2008.
- [8] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72:027104, 2005.
- [9] S. Fortunato. Community detection in graphs. http://arxiv.org/abs/0906.0612, 2009.
- [10] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc Natl Acad Sci USA*, 99(12):7821–6, 2002.
- [11] 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.
- [12] 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.
- [13] M. E. Newman. Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA, 103(23):8577–8582, 2006.
- [14] 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.
- [15] G. Palla, A.-L. Barabasi, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.
- [16] 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.
- [17] G. Palla1, I. Derényi, I. Farkas, and T. Vicsek.

Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, pages 814–818, June 2005.

- [18] W. E. Piper, M. Marrache, R. Lacroix, A. M. Richardsen, and B. D. Jones. Cohesion as a Basic Bond in Groups. *Human Relations*, 36(2):93–108, 1983.
- [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] W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.