Understanding complex networks with community-finding algorithms

Eric D. Kelsic[1]
SURF 2005 Final Report
[1] *California Institute of Technology, Pasadena, CA 91126, USA*
(Dated: November 1, 2005)

In a complex network, edges between nodes are often distributed non-uniformly, leading to the formation of hierarchical community structures. However, finding community structure in dense networks has proven to be a difficult task. We study current community-finding algorithms on a small but dense social network from the California Institute of Technology, formed on *www.thefacebook.com*. Using knowledge of the school's undergraduate Housing system, we assess the capabilities of modularity-maximizing algorithms, which we find to work decently, and single-linkage clustering, which we find to work poorly. We also propose a new agglomerative algorithm for constructing overlapping communities using local shells, and implement methods for visualizing overlap between communities.

INTRODUCTION

Many real-world systems in the sciences, technology, and society can be represented by vertices linked together by edges to form a 'complex network'. Examples include food webs, cellular and metabolic networks, neural networks, electrical power-grids, the World-Wide Web, the Internet, collaboration networks of research scientists, social networks of acquaintances and friendships, and more [1]. However, complex networks can be difficult to analyze due to their many components and substructures [2]. Real-world networks such as these often require analysis from a holistic point of view, as their collective behavior may not be understood even given the local properties of their individual parts. Accordingly, a great deal of recent research has been devoted to finding quantities that can succinctly characterize and quantify global structures and dynamics, and to testing new descriptions that might yield better representations of the relevant internal structures [3].

Knowledge about network structure can be very useful for predicting and explaining the collective behaviors of complex systems. For example, research in complex networks has led to improved methods for vaccination [2], and has helped to explain the "small world effect", wherein everyone in a network is linked to everyone else via a surprisingly short number of connections [2, 4].

In a complex network, vertices represent individual components (for instance, a person in a social network), and the edges between them represent various interactions of the components (e.g., a friendship between two people). One important means of describing a network is to find the communities to which individual vertices might belong. This has led to insights into the structures of Congress [5], biological protein networks [6], and Amazon.com shopping preferences [7]. For example, a person's social groups might be obtained from analyzing the connections between people that he/she knows. Intuitively, a community would contain a higher density of internal edges than external edges to other communities. However, methods for finding and analyzing this so-called "community structure" are typically computationally expensive, motivating the search for fast grouping algorithms that can reliably find hierarchical structure within networks [2, 8]. Additional complications arise when overlaps are allowed between communities, as community boundaries become less precise [6].

OVERVIEW

In this paper, we discuss several results obtained from analyzing social networks from the popular college website *www.thefacebook.com*. On the Facebook, applicants with valid college emails create self-descriptive profiles and create links to their friends' profiles. As an example, we analyze the network at the California Institute of Technology (Caltech), notable for its longstanding undergraduate housing system and small size (both of which make it an ideal test network). Students choose a House at the beginning of their first year and usually retain membership throughout their undergraduate education. The Housing system impacts student life enormously, both socially and academically. We can therefore justify using knowledge of the students' self-identified House affiliations as an intuitive reference point in comparing the results of two community finding algorithms: a modularity maximization algorithm [9] and a modification of single linkage clustering [10]. Facebook networks typically have a high density of connections [17], thereby preventing the application of most existing algorithms due to their prohibitively long computation time. Caltech's relatively small population helps one to examine the computational feasibility of various community finding algorithms on dense social networks. Finally, we also propose alternative visualizations in an attempt to describe the overlapping structure between communities.

COMMUNITY STRUCTURE

Community-Finding Algorithms There exist many algorithms for finding communities within complex networks [11]. Some, such as the 'betweenness' algorithms of Girvan and Newman [8, 12], in which the most relevant links *between* communities are iteratively removed, are termed 'divisive' algorithms because they divide the network into smaller subsections. 'Agglomerative' algorithms, on the other hand, form communities by joining nodes together. Examples include single-linkage clustering [10] and modularity-based algorithms [9]. The output of many algorithms can be visualized using a dendrogram (a tree), in which the order of community splits/joins is recorded by a position on a time axis, with individual nodes positioned along the other axis.

A common example used to illustrate community structure algorithms is the Zachary Karate Club, in which an internal dispute led to the schism of a karate club into the formation of two smaller clubs [13]. A plot showing the connections between members of both clubs is shown in Fig. 1. The Karate Club network is a useful test case for community-finding algorithms because we expect any calculated communities to be very similar to the actual group memberships. In Fig. 2, we show the result of a modularity-based approach (described below) in the form of a dendrogram. The success of the algorithm is apparent in the two resulting branches that reflect the actual membership of the clubs. The only exception is member 10, who is placed in the wrong community. This might be anticipated, as member 10 has only two connections, one to each of the two clubs. One would expect non-overlapping algorithms of this type to disagree primary over 'peripheral-nodes' similar to this, in which a node's connections connect it weakly to multiple communities.

Modularity-Maximizing Algorithm As a starting point, we implemented the "modularity-based" approach described in Newman [9] because of its favorable scaling with network size. A network with n nodes can be represented with an n-by-n (unweighted) 'adjacency matrix', which has a 1 in the ith row and jth column if nodes i and j are connected, and a 0 otherwise. Modularity is defined as

$$M = \sum_i (e_{ii} - a_i^2), \tag{1}$$

where e_{ij} is the fraction of edges in the network that connect communities i and j and $a_i = \sum_j e_{ij}$. Each e_{ii} represents the number of edges internal to a community, and a_i^2 represents the expected value of internal edges assuming a uniform edge distribution. Thus, modularity measures the variance in community structure from a uniformly random graph.

Initially the algorithm starts each node in its own separate community. It then conducts a greedy search by repeatedly joining the two communities that result in the greatest increase in modularity until all of the communities are joined into one. The division with greatest calculated modularity during this search is the output of the algorithm, with run-times scaling as $O(n(m+n))$, where n is the number of nodes and m the number of

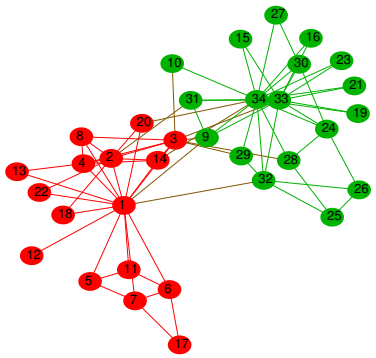


FIG. 1: Plot of the members of Zachary's Karate Club network [13], using a charge-spring embedder [14], in which nodes have been colored either red or green depending on their club affiliation. Note that some members, such as member 10, are weakly connected to both communities via a single connection.

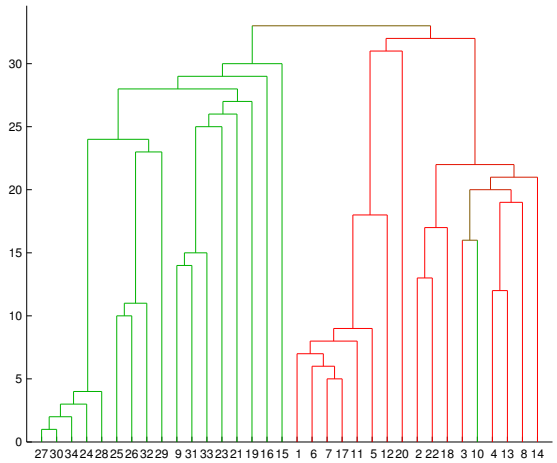


FIG. 2: Dendrogram of Zachary's Karate Club network [13] using a modularity-maximizing algorithm to group nodes into communities. Each line is colored red or green depending on the club with which a particular member aligns. Lines from nodes joined into communities are colored by the average of the community's membership. That the two main branches of the dendrogram are colored almost exclusively red or green indicates the success of the algorithm in finding the actual communities. Member 10 is the only exception, an error that might be expected due to the weak singular connections to both clubs (see Fig. 1).

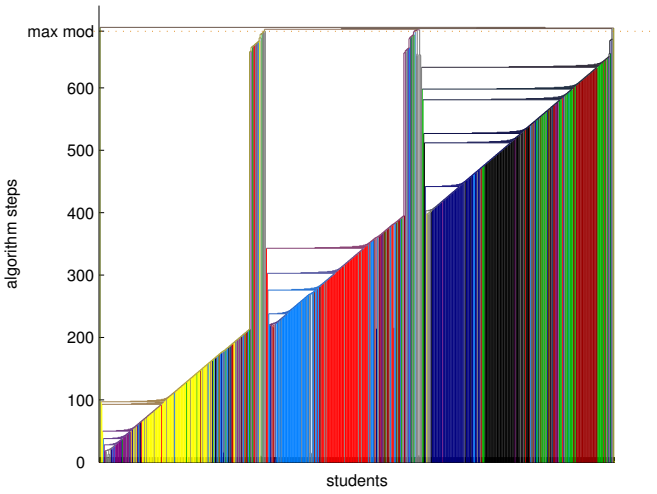


FIG. 3: Dendrogram showing the order in which nodes were added to communities with the modularity-maximizing algorithm, color-coded by self-identified House affiliation (Page: light-blue, Lloyd: yellow, Ruddock: navy-blue, Ricketts: maroon, Blacker: black, Dabney: green, Fleming: red, Avery: purple, unidentified: grey). Areas of primarily one color indicate strong communities within one House, while areas of many colors indicate a non-House-based social grouping. That many colors appear clustered horizontally demonstrates the strong House-based communities within Caltech. However, the two spiked areas that occur in the later steps of algorithm indicate that some students have friends more uniformly across Houses. Maximum modularity occurs at step 691 of 697 and is indicated by the dashed horizontal line. This results in a division with three main communities (see Fig. 4), while in fact there are eight undergraduate Houses, so the algorithm seems to overlook some important structural information. Alternative colorings of the dendrogram are also possible, and at other colleges factors such as year and major may be more important in determining social structures.

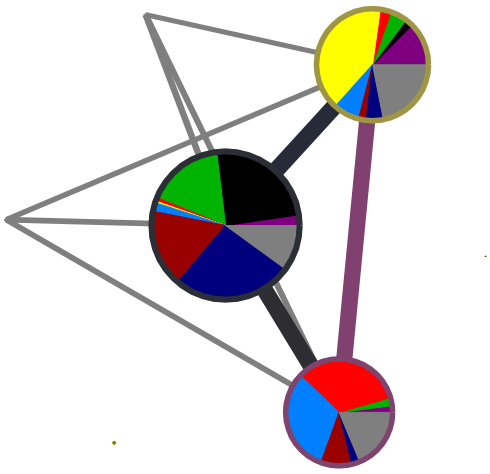


FIG. 4: Cartographic representation of the communities found at maximum local modularity using the modularity-maximizing algorithm, colored as in Fig. 1, and generated with a charge-spring embedder [14]. Community radius is determined by the number of students, and colored in radial 'pie-graph' sections according to the composition of self-identified House membership. Link widths indicate the number of connections between communities and are colored according to the weighted average House-color of students that compose the link. Grey portions of the communities are unidentified students and are most likely affiliated with the dominant House(s) in each community. Of the four smallest communities, two are shown connected to the three primary communities, while the other two are unconnected to any other nodes and are shown as very small circles. This coloring seems to show the relative friendliness of the students of each House toward each other. For example, from the Fig. one might guess that a student from Page House (light Blue) would be more likely to know more students from Fleming House (Red) than any other house, while a student in Lloyd House (yellow) might be more likely to have friends in several other Houses.

edges.

Fig. 3 shows a dendrogram that details the order in which nodes in the Caltech network were joined by the algorithm (colored by undergraduate House affiliation). The regular grouping and coloring of the dendrogram confirms our expectations that the Caltech network organizes based upon House affiliation, achieving a (reasonably high) maximum modularity of .3142 with 7 communities. Note however, that at maximum modularity (step 691 of 697) some of this structure is lost, and despite the fact that there are eight undergraduate Houses, the resulting division contains only three main groups with several tiny outlier groups. Fig. 4 shows a cartographic plot [15] of the communities at the maximized modularity. In this cartographic representation, community radius and edge width encode the number of constituent nodes and connections between communities, respectively. Communities are drawn as pie-graphs according to House composition, with edges colored by the weighted average of relevant students' House colors. It is important to note that while the House structure is discernible from the colored dendrogram, without this outside information it might be assumed that there exist only three main communities within the Caltech network. It might be expected that the modularity maximizing approach's speed comes with a tradeoff of decreased structural resolution. However, it is also possible that the increased density of the Caltech network causes the algorithm to keep joining communities past more intuitive stopping points. Perhaps an alternative definition of modularity could yield better structural divisions while still realizing the desirable quickness of the algorithm.

Single-linkage Clustering We also implement a agglom-

erative method using single-linkage clustering [10]. In this algorithm, a distance measure is defined between nodes, and each node is joined sequentially to its closest node's community. Each node begins in its own community, and joins between nodes with smaller distances are performed first. We use two similar distance metrics,

$$d_{ij} = \sum_k (n_{ik} n_{jk}) \tag{2}$$

$$d_{ij} = \sum_k \delta_{n_{ik}, n_{jk}} \tag{3}$$

Each n_{ij} is the entry in ith row and jth column of the network's adjacency matrix, and $\delta_{x_2,x_1} = 1$ if $x_2 = x_1$ and 0 otherwise. Equation (2) is essentially the dot product of the two nodes adjacency vectors; it counts the number of connections shared. On the other hand, equation (3) counts the number of connections shared or exclusively not shared. Thus $\delta_{n_{ik},n_{jk}} = 1$ if and only if $n_{ik} = n_{jk} = 1$ or $n_{ik} = n_{jk} = 0$. Fig. 5 shows the dendrograms resulting from the two distance metrics. We observe much less hierarchical structure than with the modularity-based approach. While there do appear to be small clusters which separate according to House, each House is more evenly distributed of the dendrogram, suggesting a weaker structure. The algorithm seems to have found closely connected communities on a smaller scale than that of the Houses. It is possible that this approach may be better suited to large, sparser college networks in which social groups are more widely separated from each other and where the networks' size may make other algorithms less feasible.

Overlapping Communities Most community-finding methods constrain each node to be contained within only one community, which is often unrealistic. Intuitively, many nodes in real-world networks can be grouped as a member of multiple communities. If allowed in an algorithm, this would produce communities that can contain overlapping regions. As many Caltech students are either members of multiple Houses or often socialize and collaborate with members of other Houses, it seems that an approach that realizes overlap might be insightful.

We implement a recent algorithm for finding overlapping communities described in Palla *et al.* [6]. The algorithm works by finding 'k-cliques', defined as sets in which every node is connected to at least k other nodes in the set. Increasing the value of k can be interpreted as increasing the level of connectedness within a community. Also each k-value results in a collection of communities that can include common nodes. However, testing the algorithm on the Caltech Facebook social network indicated that the clique-finding algorithm described in [6] was not computationally suitable for dense social networks [18].

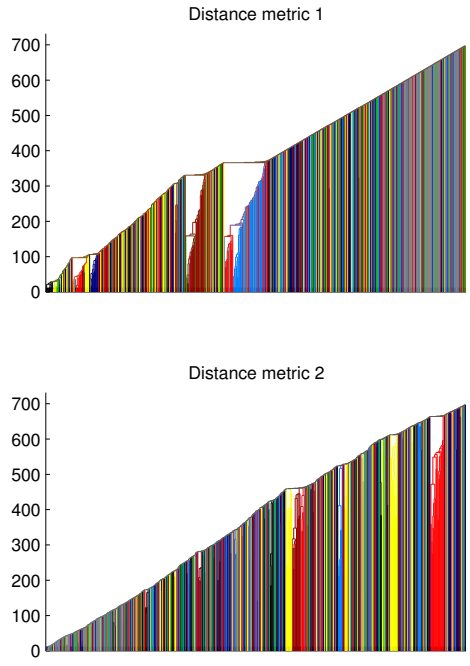
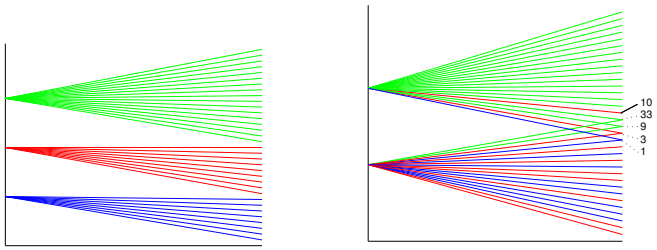


FIG. 5: Dendrograms for single-linkage clustering using the two distance metrics defined in equations (2) and (3). While some large House-based groupings appear, House members are more uniformly distributed along the dendrogram, displaying a weaker House-based organization in comparison to the modularity-based algorithm. Single-linkage clustering appears to find smaller clusters of friends with less hierarchal structure than the modularity-based approach.

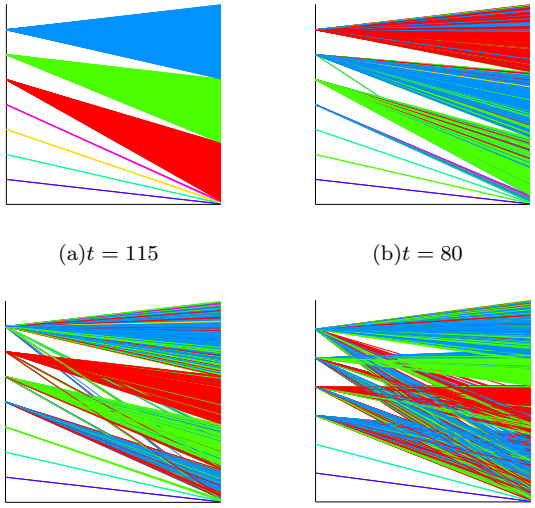
Accordingly, we propose a new method to gain insight into the overlapping structure of communities while attempting to minimize computational run-time. Recall that many agglomerative algorithms begin with each node in its own separate community, which guarantees a priori that there will be no overlap. To allow for overlap, while still retaining global network structure, we begin each community with a central node and additional nodes in its local neighborhood. A community-adjacency matrix [19] is then created that encodes the number of connections between these local communities. This matrix can be used as input for an agglomerative algorithm, such as modularity-maximization or single-linkage clustering. As a given node may be initially seeded in multiple communities, the final result of an agglomerative algorithm may now contain communities with common nodes.

There are multiple definitions that can be used to decide a local neighborhood for a given node. Conceptually, a local definition should be adjustable within some range, to allow for a variable amount of overlap. In a sparse matrix, one might consider the 'l-shell' definition



(a)$t = 10$ (b)$t = 0$

FIG. 6: Communities of the Zachary karate club at t values 10 and 0. Communities are positioned at points on the left axis (sorted in ascending size) and lines are drawn to membership nodes along the right axis. Each node is assigned a unique color based upon the community membership at $t = 10$, as shown in (a). Three communities merge into two as t is decreased to 0 in (b). Note that nodes 1, 3, 9 and 33 are contained in the intersection of both communities, demonstrating overlap. Node 10 has switched into the other community, as might be expected from our previous classification of its ambiguous connection to both clubs.



(a)$t = 115$ (b)$t = 80$

(c)$t = 50$ (d)$t = 20$

FIG. 7: Communities in the Caltech Facebook network at various t-values, drawn in the same manner as in Fig. 6. Lines are colored according to the original community membership at $t = 115$. We observe that while the three largest communities retain many of their original nodes, as the minimum value of t decreases they begin to overlap with each other and form new communities. The value of t thus controls a variable amount of overlap.

of Bagrow and Bollt [16], in which a set of nodes surrounding a central node is expanded to connected nodes until the ratio of new outgoing edges to previous outgoing edges falls below a threshold value. However, the smallest l-shell is just the set of immediately adjacent nodes, and in a dense network this leads to an almost entirely filled community-adjacency matrix, which removes most structural information. As an alternative local metric, we use the number of triangles that include both nodes to be a measure of the closeness between nodes. We can then define a 't-shell' about a central node, such that all nodes contained in a t-shell will be a member of more than t triangles that include the central node. In our algorithm, overlapping communities thus have more or less precise boundary regions, which changes with higher or lower values of t.

While we use the modularity-maximizing algorithm to combine communities here, in principle any agglomerative algorithm can be employed. To illustrate the method, Fig. 6 displays the Zachary karate club communities calculated with the maximum and minimum t values of 0 and 10 respectively. We plot communities on the left axis and nodes on the right, and draw a line between each community and its constituent nodes. Nodes are colored according to their community membership at $t = t_{max}$, in this case at $t = 10$, thus Fig. $6(a)$ contains 3 communities whose membership corresponds to the 3 branches of the dendrogram in Fig. 2. Decreasing t to 0 causes two of the communities (which correspond to the same club) to merge together as shown in $6(b)$, confirming the intuition that these two communities are similar. Four nodes are shown as belonging to both communities, indicating that these members have more connections that span the two clubs.

Fig. 7 shows four similar plots using the Caltech Facebook network, calculated for t-values of 115, 80, 50, and 20. Again, lines are drawn from communities on the left to the nodes on the right, and colored according to the community colors of nodes in Fig. $7(a)$. At $t = 115$ (the maximum t-value), each initial community consists of single central node, and these are joined into the seven non-overlapping communities described above (see Fig. 4). As the minimum value of t decreases, these three communities retain many of their nodes, although a small percentage of nodes begin to swap communities, following which we observe the formation of new communities and overlap between communities. At $t = 20$ we observe four large communities with large amounts of overlap. Thus, by varying the minimum t-value we can regulate the amount of overlap between communities.

We can plot the results over many t values as shown in Fig. 8. We assign each community a unique color based upon the non-overlapping division with $t = t_{max}$, and color each node according to its community's color. We then vary the threshold t value from this maximum to 0, and record the resulting communities. For each t value, each community is colored according to the average color of its constituent nodes. We plot the color of each node's community vertically, or, if a node be-

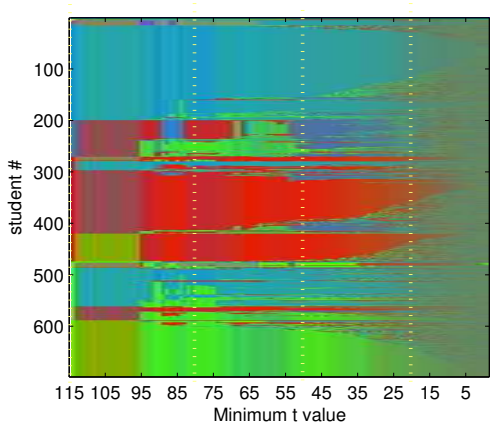


FIG. 8: Image showing the average community colors at maximum modularity, with t values ranging from 0 to 115. Unique initial node colors are assigned based upon community membership with the maximum t value, resulting in a non-overlapping structure. Nodes that belong to multiple communities are randomly assigned one community's color and node positions are sorted to preserve areas of common community membership. Noting the community colors, one can track the merging of communities at various t values, with regions of high color variance indicating overlapping regions of distinct membership. As the minimum t value decreases, communities contain more nodes and their colors thus approach a homogeneous average. For illustrative purposes, we choose four values of $t = 115$, 80, 50 and 20 (indicated by dashed vertical lines), to demonstrate our method of displaying overlapping structure, and show the resulting communities in Fig. 7.

longs to multiple communities, we randomly choose one community's color. The result shows initially distinct communities that merge and overlap as t decreases to 0.

DISCUSSION AND CONCLUSIONS

Using the Caltech Facebook network, we were able to assess the effectiveness of agglomerative techniques such as modularity maximization and single-linkage clustering. We find that maximizing a network's modularity reveals Caltech's House-based social structure, though not at maximum modularity. Subsequent work may focus on creating a modified version of modularity which can help highlight structural elements of different sizes. We also present an algorithm for calculating overlapping communities and methods for visualizing overlap in networks.

Additionally we obtained a larger Facebook network containing 100 schools of varying populations (1,000 - 55,000) and connections between schools, which we have not had time to analyze. We plan to continue researching this larger network both to assess and improve the performances of community-finding algorithms and to make comparisons of schools' social structures that may be relevant to prospective students.

ACKNOWLEDGEMENTS

We would like to thank the Caltech SURF program, and in particular special thanks to SURF mentor Mason Porter, Adam D'Angelo for providing Facebook data, Mark Newman for useful advice, and SURF donor Arthur R. Adams.

[1] Strogatz, S. H. Exploring complex networks. **2001**. *Nature* **410(6825)**, 268–276.
[2] Newman, M. E. J. The structure and function of complex networks. **2003**. *SIAM Review* **45(2)**, 167–256.
[3] Newman, M. E. J. Detecting community structure in networks. **2004**. *European Physics Journal B* **38(2)**, 321–330.
[4] Milgram, S. Small world problem. **1967**. *Psychology Today* **1(1)**, 61–67.
[5] Porter, M. A, Mucha, C. P. J, Newman, M. E. J, & Warmbrand, C. M. A network analysis of committees in the United States House of Representatives. **2005**. *Proceedings of the National Academy of Sciences* **102(20)**, 7057–7062.
[6] Palla, G, Derenyi, I, Farkas, I, & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. **2005**. *Nature* **435(7043)**, 814–818.
[7] Clauset, A, Newman, M. E. J, & Moore, C. Finding community structure in very large networks. **2004**. *Physical Review E* **70(066111)**.
[8] Girvan, M & Newman, M. E. J. Community structure in social and biological networks. **2002**. *Proc. Natl. Acad. Sci. USA* **99(12)**, 7821–7826.
[9] Newman, M. E. J. Fast algorithm for detecting community structure in networks. **2004**. *Physical Review E* **69(066133)**.
[10] Johnson, S. C. Hierarchical clustering schemes. **1967**. *Psychometrica* **32(3)**, 241–254.
[11] Danon, L, Duch, J, Arenas, A, & Diaz-Guilera, A. Community structure identification. **2005**. *unpublished* cond-mat/0505245v1.
[12] Newman, M. E. J & Girvan, M. Finding and evaluating community strucuture in networks. **2004**. *Physical Review E* **69(026113)**.
[13] Zachary, W. W. An information flow model for conflict and fission in small groups. **1977**. *Journal of Anthropological Research* **B(4)**, 452–473.
[14] Fruchterman, T. M. J & Reingold, E. M. Graph drawing by force-directed placement. **1991**. *Software - Practice and Experience* **21**, 1129–1164.
[15] Guimera, R & Amaral, L. A. N. Functional cartographic of complex metabolic networks. **2005**. *Nature* **433(7028)**, 895–900.
[16] Bagrow, J. P & Bollt, E. M. A local method for detecting communities. **2004**. *Submitted to Physics Review E, cond-mat/0412482.*
[17] The Caltech Facebook network contains 698 students, with an average 43 friends each. Caltech's adjacency matrix is thus 6.2 percent filled.
[18] Preliminary runs of the algorithm on the Caltech net-

work ran for more than a week with no sign of stopping. While networks mentioned in [6] had considerably more nodes than the Caltech network (up to 30,739 nodes for a co-authorship network as compared to 698 students at Caltech), the average node degree was considerably less (4.43 as compared to 43.0). This results in a much greater

set from which to find the algorithm's required k-cliques.

[19] Similar to an adjacency matrix, the i, jth entry for a 'community adjacency matrix' contains the number of connections between community i and community j.