Social Networks 35 (2013) 439–450

Contents lists available at SciVerse ScienceDirect

Social Networks

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

Familial groups in social networks

James Nastos*, Yong Gao

Department of Computer Science, University of British Columbia Okanagan, Kelowna, British Columbia, Canada V1V 1V7

ARTICLE INFO

Keywords:
Networks
Community
Partition
Graph theory
Familial groups

ABSTRACT

Many structural definitions for social community have been proposed in attempt to characterize and further understand the structure of social relationships. Algorithms using quantitative concepts such as centrality measures, spectral methods and other clustering measures have been used to compute social communities. While these methods have had much success in extracting meaningful subgroups in social and biological (and other) networks, they do not necessarily reveal the defining structure of social attraction.

We propose a new definition here for social community with a very clear and simple graph-theoretic structure which can also be realized as a new clique-relaxation. This structure evolved from Freeman's definition of social community, and this definition is further supported by long-standing sociometric principles such as Granovetter's *weak-tie hypothesis* or Faust's and others' studies on how global structure can be inferred from a complete understanding of local structures (although our definition goes beyond dyadic and triadic configurations). We provide computational results that show our simply-stated structural definition reveals communities that correspond almost identically to, and sometimes are better than, the widely used centrality-based methods.

We name these new communities *familial groups*, inspired by the network structures resulting from inheritance or blood-line relations. These structures form naturally in hierarchical arrangements such as in corporate settings. Using results from graph theory, our structural definition for familial groups also immediately implies a ranking of the individuals within the group, easily identifying leaders and subcommunities.

1. Introduction

Our research originates from a study of a paper of Freeman (1996) which proposed a definition to capture the essence of communal structure in social networks based only on the information of social ties between individuals. Freeman gave a definition for social community which was more general than a clique by imposing overlap conditions on maximal cliques. Falzon (2000) modified that definition and described algorithms to find such communities.

The notion of *network robustness* captures the property that desirable characteristics of the network still hold under some failures or disconnections. Various applications of robustness can be found in Newman (2010). The notion of *robust communities* has been studied in the context of networks undergoing perturbations (Karrer et al., 2008; Lemmouchi et al., 2009). Balasundaram et al. (2011) comment that the desired properties of a social community are (i) *familiarity*; (ii) *low diameter*; and (iii) *robustness*, where the removal of few individuals or connections will not break the community.

Also of some interest related to communal structure is the measure of *social position* (Kazienko and Musial, 2007) which ranks individuals within a community in a way that measures the importance of the individual in the group.

We give a formal treatment of communities comprising of sets of actors robust under the action of individuals leaving the community, and show that the resulting structures are a natural choice in defining communal structure for social networks. These communities give a natural ranking scheme for the social positions of the individuals, essentially predicting a hierarchical arrangement within the community. We note that these hierarchical arrangements vary from the usual notion of *hierarchical clustering* (Ward, 1963) in that hierarchical clustering arranges clusters in a hierarchical dendrogram while we will be interested in predicting the hierarchy of individuals in a community.

The methods in this paper can be applied to networks with or without vertex and edge weights, but we will limit our discussion to unweighted networks with undirected connections. This paper also serves to bring a body of knowledge from the study of *graph classes* (see e.g. Golumbic, 2004; Brandstädt et al., 1999) into social network analysis where it is surprisingly under-used.

* Corresponding author. Tel.: +1 250 807 9597; fax: +1 250 807 0001.
E-mail addresses: jnastos@interchange.ubc.ca (J. Nastos), yong.gao@ubc.ca (Y. Gao).

0378-8733/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.socnet.2013.05.001

Homans (1950) propositioned that a general understanding of large social networks can be inferred from fully understanding the small groups within it. Davis and Leinhardt (1967) support this claim by characterizing all possible *triadic configurations* (that is, all possible arrangements of three individuals connected with directed edges) and classifying the forbidden triads that should not appear in social groups. They empirically support this by looking at a large number of sample social group networks. Vertex degrees (density), dyad and triad distributions are still the dominant local structures used to infer global structural properties of a social network, see for example Faust (2008). This study infers global information about social networks by looking at all 4-sets of individuals in the network. Indeed, in algorithmic graph theory, the structure of every 4-set of vertices has been shown to completely characterize many important global structural properties of a graph. A prime example of this is the *semi-strong perfect graph theorem* of Reed (1987).

In Section 3, we show that the robustness criterion applied to Freeman's communities gives rise to a structural characterization which we then use to find social communities in a number of sample networks.

2. Graph-theoretical concepts and definitions

A *graph* $G = (V, E)$ is a collection V of objects (vertices), and a collection of edges E, each of which joins two vertices. We will use the terms *network* and *graph* synonymously, as well as *vertex*, *individual* and *node*. Connections between nodes will be referred to as *edges*. A *subgraph* $H = (V_H, E_H)$ of a graph $G = (V_G, E_G)$ is a graph where $V_H \subseteq V_G$ and $V_E \subseteq E_G$. A subgraph H of G is an *induced subgraph* if an edge xy exists in H if and only if the edge xy exists in G. That is, given a graph G, we can specify an induced subgraph H of a graph G by simply referring to the set of vertices that are in H, and the edges that exist among those vertices in G will be the edges of H. A *clique* is a set of nodes in which every pair of nodes is connected with an edge.

Note that if S is a set of vertices of G, the graph $G - S$ obtained from removing the vertices in S from G is an induced subgraph of G. A graph property is *hereditary* if whenever it holds for a graph G, it also holds for every induced subgraph of G. Examples of hereditary properties include planarity and being acyclic, while some examples of properties which are not hereditary are being k-connected or having diameter d.

A P_4 is a set of four vertices $\{a, b, c, d\}$ with edges ab, bc, cd, and no other edges. That is, a P_4 is a set of 4 vertices that induce a path. A C_4 is a set of four vertices $\{a, b, c, d\}$ with edges ab, bc, cd, da. A C_4 is sometimes called a *square* or a *4-cycle*. Observe that a C_4 contains a path on 4 vertices as a subgraph, but not as an induced subgraph.

A graph is called (P_4, C_4)-*free* if it does not contain any P_4 or C_4 as an induced subgraph. The class of (P_4, C_4)-free graphs is well-studied and is also known as the class of *quasi-threshold graphs*[1] (Chvátal and Hammer, 1977), *trivially perfect graphs*[2] (Golumbic, 1978), *comparability graphs of trees* (Wolk, 1962), or *arborescent comparability graphs* (Donnelly and Isaak, 1999). The property of being F-free for any fixed induced subgraph F is a hereditary graph property.

[1] The term *quasi-threshold* comes from the fact that these graphs generalize *threshold* graphs, which are graphs created from weighted vertices, and two vertices u and v are joined by an edge if and only if the sum of the weights of u and v is above a given fixed threshold value. Threshold graphs are also equivalently characterized as being the $(C_4, P_4, 2K_2)$-free graphs.

[2] A graph was defined to be *trivially perfect* if every maximal clique intersects the maximum independent set of the graph. The name derives from the fact that these are easily shown to be a subclass of *perfect graphs* which are an important class in algorithmic graph theory. It turns out that a graph is trivially perfect if and only if it is (P_4, C_4)-free.

2.1. Cohesive groups and existing techniques

This paper will solely focus on partitive methods, but the notion of overlapping communities has also been explored in various ways. A partitive method will attempt to classify each vertex into a single community while overlapping communities allow vertices and sets of vertices to exist in multiple communities simultaneously. While Freeman (1996) uses the idea of overlapping cliques to define a community, Freeman clearly stipulates that each individual is assumed to belong to a single community and that his definition of community partitions a given network. In contrast, the *rolling k-clique* definition of Palla et al. (2005) and the clique-graph definition of Evans (2010) use overlapping cliques similarly to Freeman's but only while considering cliques of a fixed size. The definitions in Palla et al. (2005) and Evans (2010) result in communities that may overlap.

We quickly survey some of the existing methods and structures used in identifying network communities.

2.1.1. Cliques as cohesive groups and its generalizations

In data where relationships (edges) between objects is expected to be transitive, the resulting graph that represents that relationship is expected to arrange itself into a disjoint union of cliques. These graphs are known as *cluster graphs* and they form a hereditary class of graphs which can alternately be characterized as the P_3-free graphs. We can thus say that this P_3-free *local structure* on three vertices completely characterizes the network structure.

When data-gathering methods are incomplete, the resulting network will be close to a cluster graph but with some missing edges. In other applications where false-positives appear, the corresponding graph will contain extraneous edges between the implied components. The common approach taken to identify the implied clusters of such networks is through graph editing: the addition or removal of as few edges as possible in order to obtain the desired structure. The associated algorithmic problem is known as *Cluster Editing*.

The idea of cluster editing is also known as *correlation clustering* (Bansal et al., 2004) in machine learning and other fields of computer science and approximation algorithms have been designed for the problems of partitioning a network to minimize the number of inter-cluster edges and/or maximize the intra-cluster edges.

As clique components are a very stringent condition to impose on a social collection, various generalizations to cliques have been explored in the literature. Early attempts include that of Luce's *n-cliques* (1950) and Alba's *n-clans* and *n-clubs* (1973) (Mokken, 1979). The computational problems of determining whether a graph contains an n-clique or n-club were shown to be NP-complete in Balasundaram et al. (2005).

Another generalization of cliques is the k-plex, which is a collection of n vertices in which every vertex is adjacent to at least $n - k$ other vertices in the collection (Seidman and Foster, 1978). Finding a k-plex of size n was shown to be NP-complete by Balasundaram et al. (2011). Cluster graphs have also been generalized in such a way that cliques – rather than being completely disconnected – may intersect in at most s vertices or t edges (Fellows et al., 2011). Such graphs admit a finite forbidden induced subgraph characterization as well: for example, the graphs in which cliques are allowed to intersect in at most one vertex are exactly the diamond-free graphs, where the diamond is the one-vertex extension of a P_3 depicted in Fig. 1. Again, the structure of this class of graphs is characterized via a characterization of its local structure.

The class of cluster graphs can be generalized through its local structure by allowing P_3s to exist in a graph, but forbidding some of the possible extensions of a P_3. A complete set of isomorphically distinct one-vertex extensions of a P_3 are given in Fig. 1.

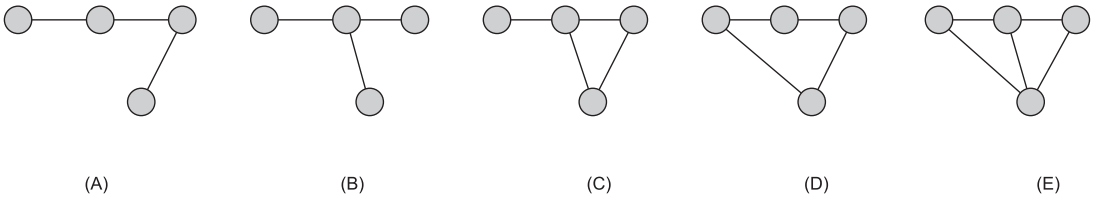

Fig. 1. The one-vertex extensions of a P_3: (A) P_4; (B) claw; (C) paw; (D) C_4; (E) diamond.

2.1.2. Graph clustering methods

Other community-finding methods exploit the fact that as much a community should be densely connected within itself, it should not be as heavily connected to the rest of the network. There are several examples of definitions that require exterior sparsity in addition to interior density. In 1969, this idea was expressed in an *LS-set* which is a set S such that every vertex in S has more neighbours in S than in $G - S$ (Luccio and Sami, 1969).

Many methods have been developed to identify dense clusters in networks. The measure of *betweenness centrality* (Freeman, 1977) has been used by Girvan and Newman (2002) to identify cohesive groups. The strategy in the Girvan–Newman algorithm is to identify and remove edges of high centrality since such edges are typically regarded as being edges that cross between separate communities. As edges are removed, the modularity of the network is measured and once all edges have been removed, the step of this process which resulted with the largest modularity score gives a natural partition of the network into groups. The process runs in polynomial time and has been shown to produce meaningful results on real networks; however, its focus on edges which are not in communities does not imply or suggest a structure of community that we are after.

Other algorithms, similar to the Girvan–Newman algorithm, have been suggested with *betweenness centrality* swapped out for another measure. For example, Radicchi et al. (2004) observe that edges that cross dense groups are not typically in as many triangles as edges inside the dense clusters. Removing edges that appear in the fewest triangles results in partitions similar to those found by the Girvan–Newman method. These methods are also shown to generalize LS-sets.

Many of the methods which use a structural definition and seek out these structures in networks result in problem formulations which are inherently NP-complete, as mentioned earlier with clique, clan, club, and k-plex finding. While this may be a computational obstruction, there are a variety of techniques that can be used to extract these desirable structures from networks. For instance, Integer LPs such as those in Balasundaram et al. (2011) offer an exact algorithm for these problems, but also lend themselves readily to faster approximation algorithms. Fixed-parameter tractability (FPT) is another technique that has been used to create efficient algorithms for NP-complete problems. Finding network clusters via cluster editing, for example, has been studied extensively in the FPT framework (Böcker et al., 2011) with great success.

The literature on finding cohesive subgroups of networks is vast, and methods such as spectral methods (White and Smyth, 2005) and probabilistic model-fitting (Hastings, 2006) have been explored. Many such methods are summarized in the surveys by Fortunato (2010) and Schaeffer (2007).

Another example of overlapping communities can be found in Mishra et al. (2008), which parameterizes the measures of internal density and external sparsity. Let α and β be two values between 0 and 1: then a set C is a (α, β)-cluster if (1) every v in C is adjacent to at least $\beta|C|$ of the vertices in C, and (2) for every u outside of C, the number of vertices in C that are adjacent to u is less than $\alpha|C|$. A desired property of this definition is that it is possible to have sets C_1 and C_2 that have a significant intersection size, are each (α, β)-clusters, while $C_1 \cup C_2$ is not a (α, β)-cluster.

We emphasize here that our method of identifying social communities is but one of many possible ways of computing clusters or even defining community, even though we believe ours is a better and more flexible one as compared to existing ones. In addition to the case studies performed in this paper (see Section 4), more empirical as well as theoretical work is needed to further validate the relevance of the proposed method to social network analysis.

3. Familial groups

We define a new community structure by generalizing cluster graphs through local structure. Rather than each community being a clique as in the case of cluster graphs, we relax that definition so that our components are family-like structures. As in the case of cluster editing where communities are found by finding a closest P_3-free graph, we find *familial groups* of a network by finding a closest (P_4, C_4)-free graph.

Definition 1. The familial groups of a network G are the connected components of a closest (P_4, C_4)-free graph.

The measure of what a "closest" network can vary, but following the paradigm of the cluster editing problem, we shall use the measure of total edge additions plus edge deletions. These are collectively referred to as *edge edits*.

Since there are several ways to "destroy" a P_4 or C_4 with edge edits, the resulting decomposition may not be unique (but this is a reality in P_3-editing for correlation clustering, and many other community-finding methods as well.) An example of two different outcomes of editing a given graph with an equal number of edge edits is shown in Fig. 2. Under a framework of weighted modifications, for instance, a cost of α for adding an edge and β for deleting an edge, one could weigh one decomposition better than another. For instance, if we were interested more in seeing how a network decomposes into groups, we would set $\alpha > \beta$, and vice versa if we were more interested in seeing how individuals in a community are organized. Weights on individual edges with perturbations can ensure a unique optimal decomposition into familial groups if desired. In this paper, we only consider $\alpha = \beta = 1$ and we

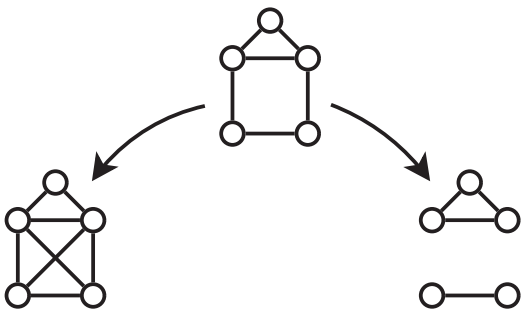

Fig. 2. Two equally weighted outcomes of modifying a graph to a closest (P_4, C_4)-free graph.

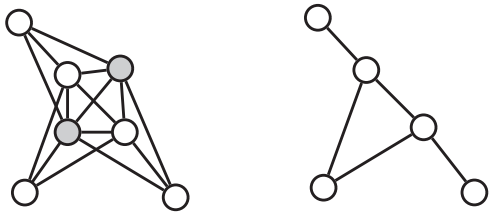

Fig. 3. A community satisfying Freeman's transitively overlapping clique condition (left) that is not hereditary. Removing the two filled vertices yields a graph (right) which no longer satisfies the definition.

are interested in modifying networks using as few modifications as possible.

3.1. Properties of familial groups

In the following subsections, we give support to the idea that a quasi-threshold graph (a (P_4, C_4)-free graph) is an *ideal structure* for networks composed of family-like or hierarchically organized communities.

3.1.1. Hierarchical representation of familial groups

There are several characterizations for (P_4, C_4)-free graphs, many of which can be found in Yan et al. (1996) where they are called *quasi-threshold graphs*. Along with the forbidden induced subgraph characterization, we will be interested in the rooted tree representation of quasi-threshold graphs.

Every quasi-threshold graph G can be arranged into a forest-like structure (a set of tree-like structures) in which every vertex is adjacent (in G) to every descendant in the tree. In particular, the root of a tree T is adjacent (in G) to every vertex in T, and there does not exist an edge joining two vertices in separate trees. An example of a quasi-threshold graph and its associated comparability tree are given in Fig. 4. Note that every leaf in a tree is adjacent to all of its ancestors and that every set of vertices along a root-to-leaf path forms a maximal clique of the graph.

In the graph in Fig. 4, vertex 5 is a universal vertex (a vertex adjacent to every vertex). It is the root of the associated tree. The rest of the vertices form two connected components: {4} and {1, 2, 3, 6}. In the component {1, 2, 3, 6}, vertex 2 is universal and is the root of the subtree consisting of vertices 1,2,3, and 6. The other subtree consists of a single vertex 4. In the subtree rooted at 2, the positions of 1 and 6 can be interchanged with each other arbitrarily because they are structurally equivalent. In this example, the whole network is a familial group. If vertex 5 is removed from the network, then we will have two separated smaller familial groups {1, 2, 3, 6} and {4}. In turn, after removing vertex 2 in the group {1, 2, 3, 6}, we will get two even smaller familial groups {1, 6} and {3}.

Quasi-threshold graphs are natural structures that arise from modeling certain applications. For instance, if a graph is created on a set of species such that an edge is drawn between two species if and only if they have an ancestor/descendant relationship, then the graph created will form a quasi-threshold graph if the information obtained was error-free. Another example is that of a corporate structure in which every employee (except one) has a direct supervisor, and that commands can be passed to an employee from her supervisor or her supervisor's supervisor, etc. When an edge is joined between any two individuals on which a command can pass, the resulting graph is a quasi-threshold graph.

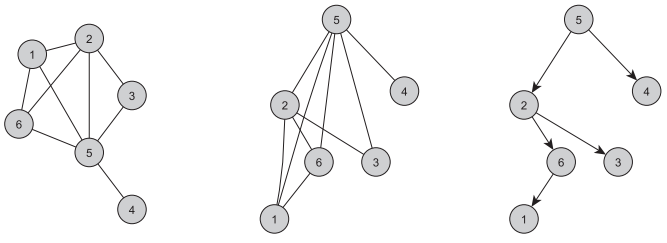

Fig. 4. (Left) A quasi-threshold graph; (center) a tree arrangement of the graph on the left; (right) the comparability tree of the quasi-threshold graph on the left.

3.1.2. Familial groups as robust communities

Freeman gave a definition for social community which uses the idea of *clique overlaps*. Two cliques overlap if they intersect in at least one vertex. The definition (Freeman, 1978) can be summarized as follows: a set of maximal cliques $C_1, C_2, C_3, \ldots, C_k$ which induces a connected graph forms a community if the cliques C_i overlap transitively. That is, for any three cliques C_i, C_j, C_k, if C_i overlaps C_j and C_j overlaps C_k, then C_i and C_k must also overlap. Freeman rationalized his definition by stating that an individual should be contained in a single community (that is, a network should decompose into disjoint communities), that it generalized cliques, and that it is applicable to networks in which only relationships (of unknown strength) between pairs of individuals was known. That is, his definition applies to undirected and unweighted graphs.

We will enforce a level of robustness to this definition of community to create a tighter definition of community. The removal of any vertices from a graph G leaves behind a graph H which is an induced subgraph of G. The robustness we impose can be stated as follows: a set of vertices S will form a familial group if *S and every connected induced subgraph of S* satisfies the above transitively overlapping clique property. Socially speaking, the community remains intact if the removal of any number of individuals leaves the group connected. Or, in the case that some "important" individuals leave the community and disconnect it, then the remaining connected components will themselves form smaller communities. An example of a community which satisfies Freeman's transitively overlapping clique property, but not hereditarily, is depicted in Fig. 3.

We show that simply requiring Freeman's transitively overlapping clique condition to be hereditary yields a formulation of social community which exactly corresponds to connected (P_4, C_4)-free graphs.

Theorem 3.2. *A connected set S of vertices satisfies Freeman's transitively overlapping clique condition in every connected induced subgraph if and only if S induces a connected (P_4, C_4)-free graph.*

Proof. If S satisfies the transitively overlapping clique condition for every induced subgraph, then it cannot contain an induced path on 4 vertices ab, bc, cd since each edge is a maximal clique while ab overlaps with bc and bc overlaps with cd, but ab does not overlap with cd. Similarly, it cannot contain an induced cycle on 4 vertices ab, bc, cd, da for the same reason. So any graph satisfying the transitively overlapping clique condition must be (P_4, C_4)-free.

Conversely, if a connected graph S is (P_4, C_4)-free, it must have a vertex u which is adjacent to all other vertices in S (Yan et al., 1996). Since there is such a *universal* vertex u in every connected component of a (P_4, C_4)-free graph, every maximal clique in a connected component must include u, and so all maximal cliques in the connected component overlap, at least on vertex u. Consequently, the cliques overlap transitively. □

3.1.3. Familial groups as an extension of triadic closure

Some sociometric data not only measures when two objects are related, but also measures the strength of the tie between them. In 1973, Granovetter formulated as follows.

The Weak-Tie Hypothesis: if a is strongly tied to b and a is strongly tied to c, then it is more likely than not that b and c are at least weakly tied to each other.

Granovetter observed that the weak-tie hypothesis can be used to assert that the most unlikely triad to appear in a social group is when a is strongly tied to b and a is strongly tied to c, while b and c have no social relation between them. He goes on to propose a graph edit operation called *triadic closure* which adds at least a weak tie between b and c. In the framework of unweighted edges, this is exactly the condition that a social group is P_3-free as discussed previously.

We generalize the forbidden restriction on triads to forbidding certain configurations on 4 nodes, the P_4 (which contains two induced P_3s) and the C_4 (which contains four induced P_3s).

An intuitive argument further supports the restriction of long induced paths or cycles from social communities. A close-knit community should have relatively low diameter, and the existence of two vertices of geodesic distance d from each other would imply the existence of an induced path of length d. Thus a social community should be P_d-free from some relatively small value of d. An argument against the existence of induced 4-cycles in communities is that if a is tied to b, b to c, c to d, then d tied to a, it is highly likely that a will get to know c or b to know d. That is, ac and bd are highly likely chords in the cycle $abcd$. As such, it is reasonable to expect social communities to be P_d-free and C_4-free for relatively small values of d.

While we will be concerned with (P_4, C_4)-free graphs here, larger and more relaxed communities could be identified if the focus is changed to (P_5, C_5)-free graphs or (P_5, C_4, C_5)-free graphs.

3.2. Hardness of finding familial groups

The computational problem of interest here is as follows.

Problem 1. Quasi-Threshold Editing(G, k): *Given a graph G and an integer k, is there a set S of edge deletions and a set T of edge additions such that $|S| + |T| \leq k$ and $G - S + T$ is (P_4, C_4)-free*?

The quasi-threshold edge-addition problem has been studied in Guo (2007) and the quasi-threshold edge-deletion problem in Nastos and Gao (2010). To our knowledge, the Quasi-Threshold Editing problem has not yet been studied directly.

Given any graph as input, the algorithm of Chu (2008) decides in linear time $(O(m+n))$ whether the input is quasi-threshold and in the case that it is not, a P_4 or a C_4 will be produced. The computational status of the problem of finding the closest quasi-threshold graph (in terms of the number of edge modifications) was stated as an open problem in Burzyn et al. (2006), Mancini (2008) and again in Liu et al. (2011). We resolve this open question by showing that this problem is NP-complete by observing some extensions on a theorem in Liu et al. (2011).

Theorem 3.3. *Quasi-Threshold Editing is NP-complete.*

Proof. We outline the proof idea here. For a complete description of the proof details and problem definitions, please refer to the appendix.

El-Mallah and Colbourn (1988) proved that Cograph Deletion is NP-Complete by a reduction from Exact 3-Cover. Liu et al. (2011) used the same construction to show that Cograph Editing is NP-Complete by strengthening the proof for Cograph Deletion.

A quick proof, without the details, is as follows: the reduction from Exact 3-Cover used by Liu et al. (2011) to show that Cograph Editing is NP-complete constructs a graph G^* which is also C_4-free. The optimal edge-edit set for G^* that destroys all P_4s does not produce any C_4.

Since every quasi-threshold graph is a cograph, the number of edits required to the closest quasi-threshold graph is at least the number of edits required to obtain the closest cograph.

An algorithm solving Quasi-Threshold Editing, applied to G^*, would destroy the P_4s (and not have any C_4s to worry about, as observed above) and would thus provide a solution to the instance of Exact 3-Cover. □

3.3. Algorithms for familial groups

From the finite forbidden induced subgraph characterization of quasi-threshold graphs, the problem of modifying a graph to a closest quasi-threshold graph is *fixed-parameter tractable* when using either edge additions or deletions or both (Cai, 1996). The trivial algorithm for Quasi-Threshold Editing considers all possibilities of adding/deleting an edge between each pair of vertices in a forbidden P_4 or C_4, and so finding a closest quasi threshold graph with k edits runs in $O^*(6^k)$-time, where the notation $O^*(f(n, k))$ means $O(f(n, k)p(k, n))$ for some polynomial p.

The similar problem of modifying a graph to a quasi-threshold graph using only edge deletions has been studied previously.

Theorem 3.4. *(Nastos and Gao, 2010, 2012) The quasi-threshold edge deletion problem can be solved by an algorithm running in $O^*(2.45^k)$ time.*

The methods used for algorithm improvement in Nastos and Gao (2012) for deletion problems were extended by Liu et al. (2011) to improve the trivial runtime of $O^*(6^k)$ for cograph editing to $O^*(4.612^k)$. We believe the Quasi-Threshold Editing problem can also be improved from the trivial $O^*(6^k)$-time algorithm using a similar method.

For computational feasibility, we combined the above bounded search tree method with greedy edge-edit choices according to the measure of counting the total number of induced P_4s and C_4s in the graph. By testing every possible edge-addition and every possible edge-deletion, we (greedily) chose the edge edit that resulted in the largest improvement (that is, the largest decrease) in the total number of induced P_4s +number of induced C_4s in the graph. Greedy choices were made until the brute force algorithm was able to execute on the modified graph within reasonable time.

In the next section of this paper, we analyze a selection of social networks by computing an approximate closest quasi-threshold graph with this combined search and greedy heuristic method.

3.4. Intra-communal ranking

The importance of individuals to a network or a subnetwork is often measured by means of various vertex centrality metrics. These range from simple local properties such as vertex degree to global properties such as betweenness centrality.

The actors in a connected component of a quasi-threshold network naturally arrange themselves in a rooted tree representation. This correspondence can be used to extract an importance measure of each actor within the community. Intuitively, the root or top-most vertex of a familial group is the most important node and the others are ranked by virtue of the fact that each node can be regarded as the root of a subtree. The size of a subtree under an individual will be the relevant measure of importance here, rather than a metric such as vertex degree.

For instance, in the quasi-threshold community in Fig. 5, vertex 6 has degree 5 and "oversees" 2 others, while vertex 3 has a lower degree of 4 but oversees 3 others. We perceive vertex 3 to have a more important role than vertex 6 in this community.

Hence, we define the *intra-communal rank* of a vertex v in a quasi-threshold community to be the number of vertices beneath v in the corresponding comparability tree. In the case that M vertices are structurally equivalent within the community in such a way that these M vertices all oversee d vertices beneath them in any associated comparability tree, then these M vertices can be given an intra-communal importance score of $d + (M-1)/2$.

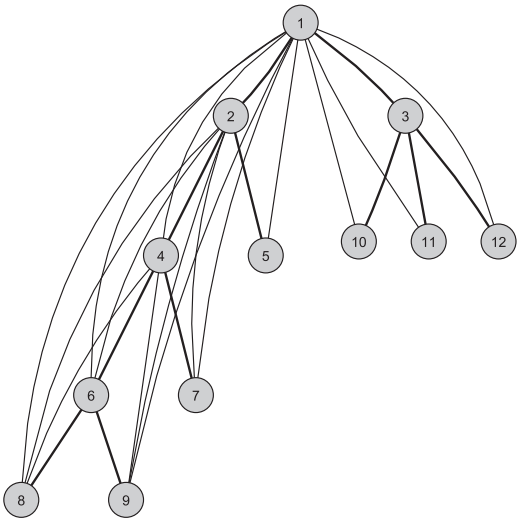

Fig. 5. The degree of an actor does not determine its social rank.

This quantitative measure of intra-communal rank is merely one way to assign a value to a vertex that captures how important it is in its familial group. There are many possible ways such a measure could be defined, and an appropriate quantitative function is perhaps a topic for future research.

4. Case studies

We present some example networks from the literature and the implied communities from a close quasi-threshold graph we computed.

4.1. Zachary's karate club

Zachary (1977) studied the social relationships between individuals in a university karate club. The club suffered a division which split the club into two, and Zachary observed that the split very closely corresponded to a min-cut that separates the two opposing individuals of largest influence.

The method of Girvan and Newman (2002) for hierarchical clustering predicts roughly the same partition that Zachary observed after the karate club experienced its social fission, with the exception of vertex 3 being misclassified. The familial groups of the karate network identified by our approach are depicted in Fig. 6 (right), where dashed lines represent edges from the network that were deleted and bold dashed lines represent new edges added in order to find the closest quasi-threshold graph. The obtained quasi-threshold partition groups the network into two groups, equivalent to the first two groups produced by the Girvan–Newman method.

An interesting result of the tree structures revealed by our approach for the two groups is that it predicts exactly two distinct components with roots of vertices 1 and 34, while Zachary's method had begun with knowing that 1 and 34 are the conflicting leaders and found a minimum cut that separated 1 and 34. Subcommunities of the two major communities can be identified as subtrees of of quasi-threshold tree. Consider the removal of vertex 1: this leaves subtrees of {12}, {5, 6, 7, 11, 17}, {2, 3, 4, 13, 14, 18, 20, 22}, which imply overlapping subcommunities when vertex 1 is regarded as a member of each of these subcommunities. We observe the similarity in the results implied by the dendrogram of Girvan and Newman, identifying a second-level community of {5, 6, 7, 11, 17} as well, and vertex 12 quickly being separated from the remaining network. The larger of these three further decomposes into overlapping subcommunities when looking under vertex 2, and these communities are {1, 2, 18}, {1, 2, 22}, {1, 2, 20}, and {1, 2, 3, 4, 8, 13, 14}.

4.2. Communities in the Les Misérables network and character importance

Les Misérables is a 19th-century novel by Victor Hugo containing 5 parts (or volumes) broken into 70 chapters. A network of 77 major and minor characters in the novel was constructed in Knuth (1993) by joining two individuals with an edge if they exist in a chapter together.

Fig. 7 shows the network and the familial groups found are distinguished by node shape and shading. The quasi-threshold graph obtained, after 64 edge edits, is shown on the right-side of Fig. 7, consisting of three large nontrivial components and several smaller ones. The predicted leaders (roots) of these three components *Jean Valjean, Marius Pontmercy* and *Fantine* with implied intra-communal scores 27, 19 and 10 (respectively), are key characters in the novel as is witnessed by the fact that their names are titles to 3 of the 5 volumes. This quasi-threshold graph correctly isolates only minor characters into trivial groups.

4.3. Lusseau's dolphin network

Lusseau (2003) studied a population of dolphins over a period of 7 years, building the social network depicted in Fig. 8 by joining an edge between two dolphins if they were observed together significantly more often than was statistically expected. The community structure of this network was studied in Newman and Girvan (2004), where the main community was identified as predominantly female and the male community split into two upon a temporary disappearance of several individuals.

Using 75 edge edits, our closest quasi-threshold graph found for the dolphin network is shown in Fig. 8 (right). Our familial grouping supports the observed communities: it shows three main groupings, one of which is almost entirely female while each of the other

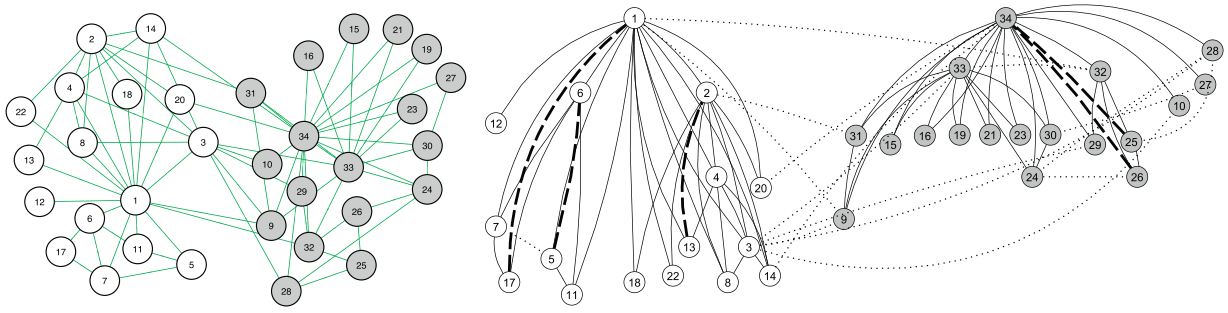

Fig. 6. (Left) Zachary's karate club network; (right) the quasi-threshold graph obtained after 20 edits.

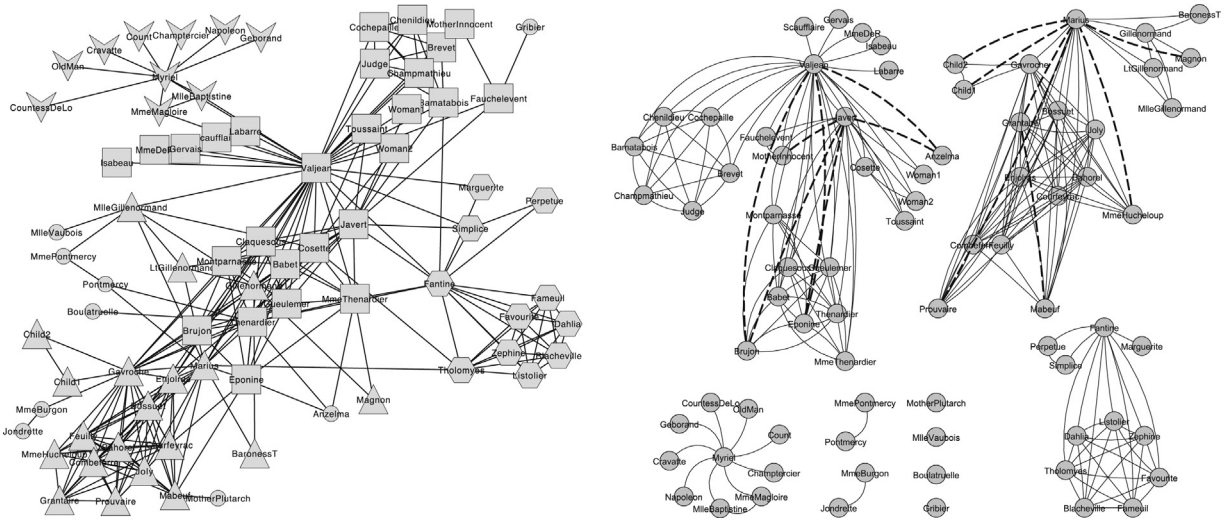

Fig. 7. (Left) Characters in the novel *Les Misérables*. The network is drawn in Cytoscape's spring embedding, while the approximated familial groups are distinguished by vertex shape and shading. (Right) The familial groups found in the left network. The bold dashed lines represent edge additions, and deleted connections are not shown.

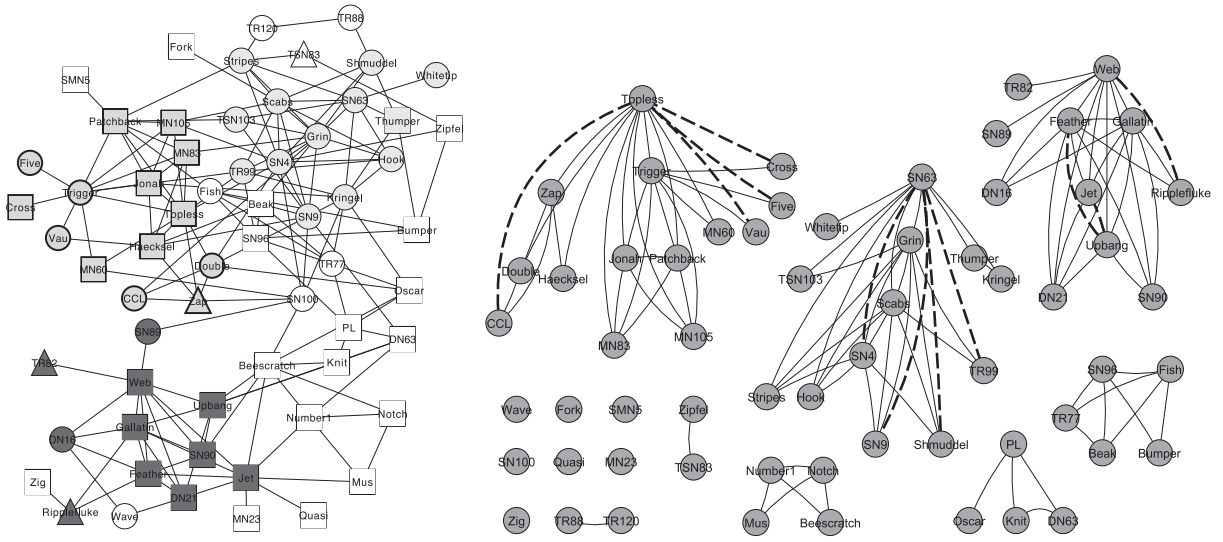

Fig. 8. (Left) Network of dolphin associations. Females are circle-shaped, males are square-shapes, and three individuals of unknown gender are depicted as triangles. The three main familial groups found are shown with light shading, light shading with thick outline, and dark shading. (Right) The corresponding comparability trees of the quasi-threshold communities found in the dolphin network.

two are mostly male. The remaining 15 dolphins are shown in the network are white nodes.

4.4. Grassland species

The left-side network of Fig. 9 is a network of grassland species interactions built in Dawah et al. (1995), and its hierarchical community structure was analyzed in Clauset et al. (2008). The network contains 1007 induced obstructions (P_4s or C_4s) and we produce a quasi-threshold graph that is 34 edge edits away from it, depicted on the right-side of Fig. 9. Each node corresponds to a type of organism such as plants (circle-shaped nodes), plant-eating organisms (square-shaped nodes) and parasitic organisms (the rest of the nodes) (Fig. 10).

Interestingly, the root node of every non-trivial familial group was found to be a herbivore. It was found in Clauset et al. (2008) that several sets of parasites were grouped together not because they fed on each other but instead because they all fed on the same herbivore. Our familial groups strongly show that the herbivores play central roles in the organization of these species.

4.5. College football network

Girvan and Newman (2002) give a network joining two American college teams together if they played against each other during the year 2000 football season. Evans writes that the data is likely the 2001 season (2010), and corrects some of the conference assignments in the data.[3] In that football season, the 115 teams are grouped into 11 conferences, with a 12th group of independent teams. Teams are usually matched against their conference-mates, an average of about 7 games against teams within their own conference and 4 games outside of conference. Girvan and Newman

[3] We thank one of the referees for bringing to our attention the corrected conference assignment made by Evans.

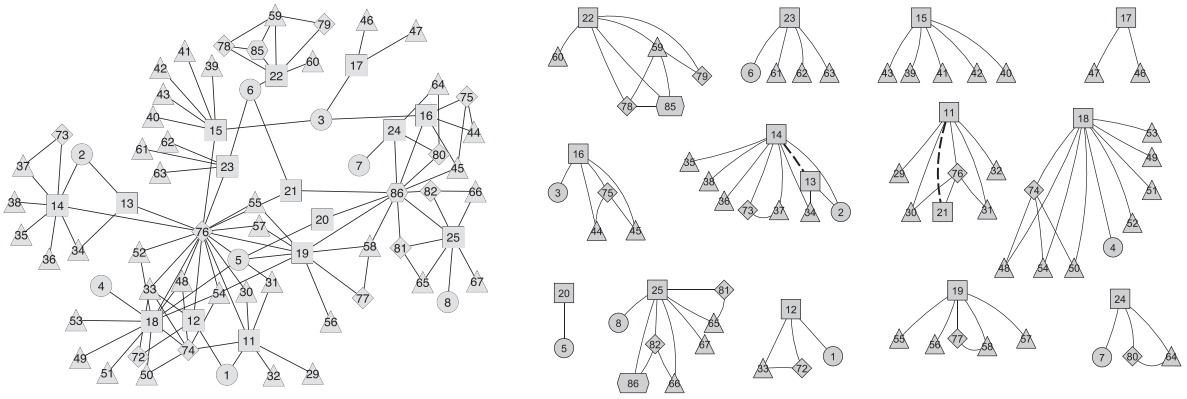

Fig. 9. (Left) Network of grassland species. Node categories are: plant (cirlce), herbivore (square), parasitoid (triangle), hyper-parasitoid (diamond), and hyper-hyper-parasitoid (hexagon). (Right) The corresponding familial groups found after 34 edge edits. Bold dashed lines are edge additions and deleted edges are not shown.

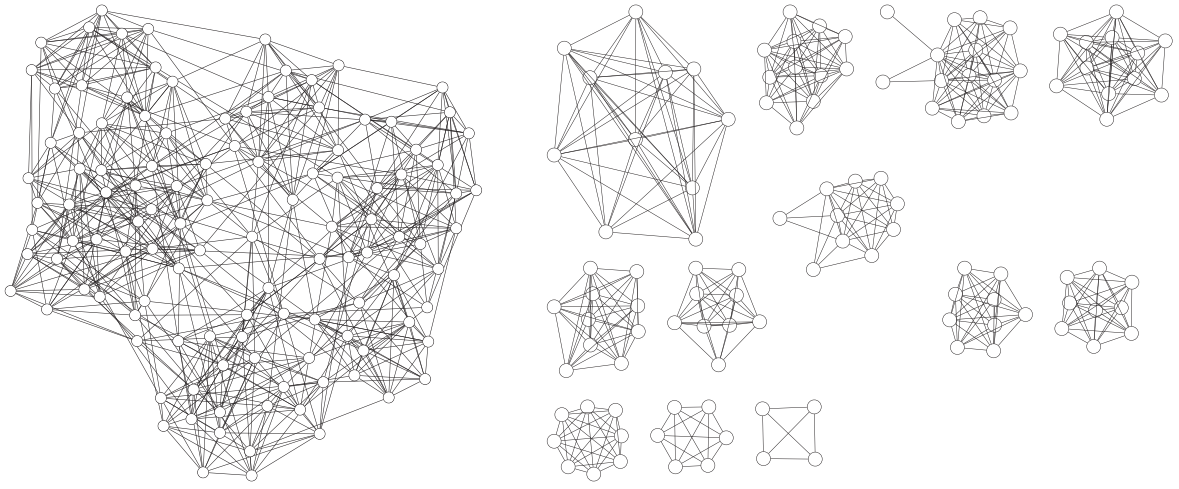

Fig. 10. (Left) The football network drawn with yEd's organic layout; (right) the corresponding familial groups found after 255 edge edits. Interestingly, exactly 12 components were found, mostly corresponding to the 12 conferences that partitioned the teams.

extracted the community structure of this network and found a near-match to the expected partitions defined by conferences.

The network began with 613 edges, and our greedy method made 255 edge edits on the network to arrive at a (P_4, C_4)-free graph.

We surprisingly found exactly 12 connected components, almost perfectly matching the 12 conference groups as labeled by Evans. A table of the familial groups found is given in Table 1. The numeric groupings in the table correspond to the connected components found. The left and right icons in each row describe which conference that team is assigned to as given by the Newman dataset and the Evans dataset, respectively.

Fig. 11 illustrates the intra-communal ranking of the teams in group 6, according to the discovered structure. The large score of Akron (the root) suggests that group 6 corresponds to the conference containing Akron, which is the *Mid American* conference. The relatively high score of Buffalo is a strong suggestion that Buffalo belongs to the same conference as Akron. The very low scores of 0 for Central Florida and Connecticut tell us that although the structure of the scheduling that year seems to associate those two teams with the *Mid American* conference, these associations are very weak, even weaker than the ties for the other leaf-node teams of larger depth. The four teams ranked with 8.5 (Toledo, West Michigan, Miami Ohio, Central Michigan) were found to be equivalently structured in the community and so the placement order of

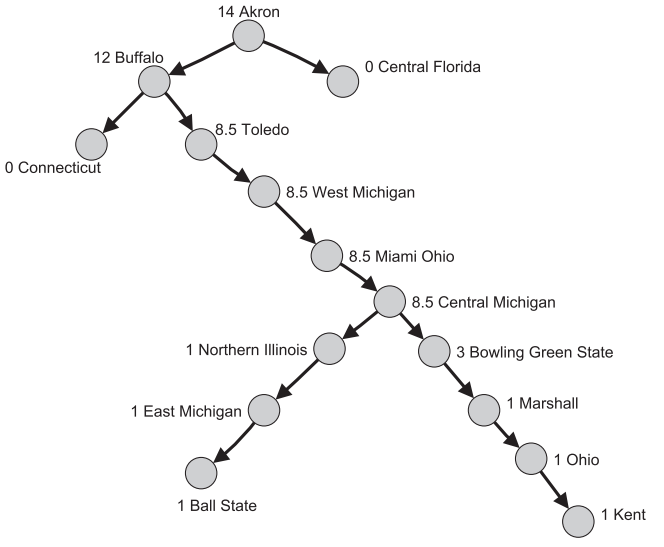

Fig. 11. The implied comparability tree corresponding to one particular familial group. The intra-communal ranking is also given for each team. This is group 6 in Table 1.

Table 1
The 12 connected components found after greedily editing-out the P_4s and C_4s from the football network. The left symbol is the conference assignment as given in Girvan and Newman's dataset, while the right symbol corresponds to Evans' corrected conference assignment. The grouping found corresponds almost exactly to the 12 conference groups described by Evans. The conference labels are depicted by the legend below.

Familial Groups Found in the Football Network								
1	○ ○	Clemson	5	□ □	Alabama Birmingham	9	◁ ⊕	North Texas
	○ ○	Wake Forest		□ □	East Carolina		■ ⊕	Utah State
	○ ○	Maryland		□ □	Houston		◁ ⊕	Arkansas State
	○ ○	North Carolina State		□ □	Louisville		◀ ⊕	Boise State
	○ ○	Florida State		□ □	Memphis		◁ ⊕	Idaho
	○ ○	Virginia		□ □	Southern Mississippi		◁ ⊕	New Mexico State
	○ ○	Georgia Tech		□ □	Tulane	10	♡ ♡	Arkansas
	○ ○	Duke		□ □	Army		♡ ♡	Auburn
	○ ○	North Carolina		□ □	Cincinnati		♡ ♡	Alabama
2	● ●	Miami Florida	6	◇ ◇	Marshall		♡ ♡	Florida
	● ●	Virginia Tech		◇ ◇	Northern Illinois		♡ ♡	Kentucky
	● ●	Boston College		◇ ◇	Western Michigan		♡ ♡	Vanderbilt
	● ●	West Virginia		◇ ◇	Akron		♡ ♡	Mississippi State
	● ●	Syracuse		◇ ◇	Ball State		♡ ♡	South Carolina
	● ●	Pittsburgh		◇ ◇	Bowling Green State		♡ ♡	Tennessee
	● ●	Temple		◇ ◇	Buffalo		♡ ♡	Mississippi
	● ●	Rutgers		◇ ◇	Central Michigan		♡ ♡	Georgia
	■ ■	Navy		◇ ◇	East Michigan		♡ ♡	Louisiana State
	■ ■	Notre Dame		◇ ◇	Kent	11	◀ ◀	Hawaii
3	▷ ▷	Michigan State		◇ ◇	Miami Ohio		□ ◀	Texas Christian
	▷ ▷	Indiana		◇ ◇	Ohio		◀ ◀	Fresno State
	▷ ▷	Northwestern		◇ ◇	Toledo		◀ ◀	Rice
	▷ ▷	Wisconsin		■ ■	Central Florida		◀ ◀	Southern Methodist
	▷ ▷	Michigan		■ ■	Connecticut		◀ ◀	Nevada
	▷ ▷	Iowa	7	⋆ ⋆	Oregon State		◀ ◀	San Jose State
	▷ ▷	Purdue		⋆ ⋆	Arizona State		◀ ◀	Texas El Paso
	▷ ▷	Ohio State		⋆ ⋆	California		◀ ◀	Tulsa
	▷ ▷	Minnesota		⋆ ⋆	UCLA	12	◀ ■	Louisiana Tech
	▷ ▷	Illinois		⋆ ⋆	Arizona		◁ ■	Louisiana Monroe
	▷ ▷	Penn State		⋆ ⋆	Washington		◁ ■	Middle Tennessee State
4	▶ ▶	Missouri		⋆ ⋆	Oregon		◁ ■	Louisiana Lafayette
	▶ ▶	Oklahoma State		⋆ ⋆	Stanford			
	▶ ▶	Baylor		⋆ ⋆	Washington State			
	▶ ▶	Colorado		⋆ ⋆	Southern California			
	▶ ▶	Kansas State	8	♣ ♣	Nevada Las Vegas			
	▶ ▶	Kansas		♣ ♣	San Diego State			
	▶ ▶	Texas Tech		♣ ♣	Wyoming			
	▶ ▶	Iowa State		♣ ♣	Brigham Young			
	▶ ▶	Nebraska		♣ ♣	Utah			
	▶ ▶	Texas A&M		♣ ♣	Colorado State			
	▶ ▶	Oklahoma		♣ ♣	New Mexico			
	▶ ▶	Texas		♣ ♣	Air Force			

○ Atlantic Coast	□ Conference USA	⋆ Pac 10	● Big East
■ IA Independents	♡ SEC	▷ Big 10	◇ Mid American
◁ Sunbelt	▶ Big 12	♣ Mountain West	◀ Western Athletic
⊕ Big West			

those 4 teams in the comparability tree is arbitrary amongst each other. Similarly, Marshall, Ohio and Kent were found to be structurally equivalent, as were Northern Illinois, East Michigan and Ball State.

To illustrate how the scores are determined, Buffalo scores 12 because there are exactly 12 nodes below it in this tree (and in every tree obtained by permuting the order of any of the structurally equivalent nodes). Toledo, however, is in an equivalence group {Toledo, West Michigan, Miami Ohio, Central Michigan}, and this group directly oversees 7 nodes below them. The intra-communal score of $d + (M - 1)/2$ with $d = 7$ and $M = 4$ gives each of these 4 teams a score of 8.5.

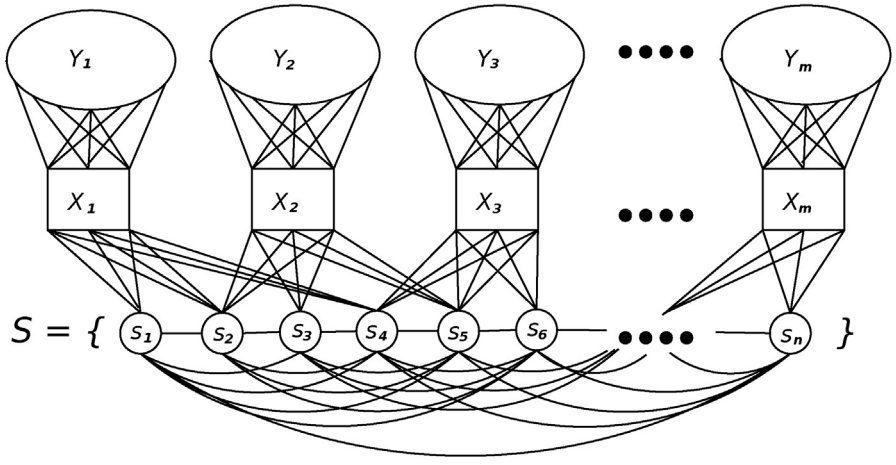

Fig. 12. An instance of Quasi-Threshold Editing when reduced from an instance of Exact-3-Cover having $S_1 = \{s_1, s_2, s_4\}$, $S_2 = \{s_2, s_3, s_5\}$ and $S_3 = \{s_4, s_5, s_6\}$.

5. Conclusions and future considerations

5.1. Summary

The main contribution of this paper is a new definition for community structure based on the graph-theoretical concept of forbidding small induced subgraphs. Furthermore, we give evidence that editing-out P_4s and C_4s in order to obtain a quasi-threshold graph yields meaningful clusterings in real networks. Due to their relatedness to family-like structures, we call these types of communities *familial groups*.

Familial groups are a natural and meaningful relaxation of many standard structures widely used in social community definitions such as cliques and k-plexes and their generalizations. They are robust against the removal of network nodes, a characterization that is not guaranteed in most other definitions of a social community.

Familial groups automatically provide a ranking of the individuals within a group and a hierarchical arrangement of a group itself – a unique feature that, to our best knowledge, no other existing model of community structure provides. These communities also retain many of the properties that a highly connected group should have such as having a low diameter. In fact, the diameter of a familial group is at most 2.

The notion of familial groups presented here can easily be modified to handle more network information, including weighted nodes and edges, directed edges, and weighted edge-edit operations. These generalizations are some of the possible lines of future investigation that still need to be taken in determining the appropriate place familial groups have in social network analysis.

5.2. Future work

Despite the theoretical and computational justifications given for the use of familial groups, there is still much work to be done in determining when this method of network clustering is desirable over other existing methods. For instance, Ravasz and Barabási (2003) found that a scale-free topology of complex networks gives evidence of hierarchically organized nodes, while networks without such structure (such as those deriving from geographical data) are not hierarchical. It would be interesting to see if the method of familial groups would be consistent with their findings by yielding meaningful results only in scale-free networks.

An aspect of (P_4, C_4) editing is that the edit set is not unique, and we can only speculate at this point how varied the found communities would be in the space of equally weighted edit solutions. A specific question we can ask is how one could define the intra-communal rank of individuals differently so that the importance of a vertex is perhaps measurable in the original network and not on the found edited graph, which is not unique. Along this line of reasoning, we wonder if there may be an easy way to predict the individuals which will end up as leaders of groups after editing to a quasi-threshold graph.

Another possible future study is to analyze how much larger the found communities become when relaxing the P_4 restriction to a P_5 restriction, and similarly relaxing to the C_4 to a C_5. Many graph classes defined by these forbidden induced subgraphs have already been studied, mostly in terms of their structure, but not necessarily in the context of modifying to such graphs. For example, while every connected (P_4, C_4)-free graph has at least one vertex such that every other vertex in the component is adjacent to it, it is also known that every connected (P_5, C_5)-free graph has a clique in it such that every other vertex is adjacent to some vertex of that clique (Cozzens and Kelleher, 1990). As far as we know, a graph modification problem (via edge deletions or edits) has not yet been studied for the class of (P_5, C_5)-free graphs. There are many other possible graph classes can that can serve as relaxations to P_4 and C_4-freeness as well.

The computational problem of editing a graph to a nearest (P_4, C_4)-free graph is still rather new. We showed here that it is in fact NP-complete, but this does not rule out fast approximation algorithms or integer linear programming formulations. Even improved exponential-time exact algorithms would be of interest, especially kernelization techniques which could reduce the size of large problem instances.

There have been many studies on inferring global structure from local analysis. With our definition of forbidding certain 4-vertex graphs, this opens the door to new structural analysis possibilities, such as probabilistic modeling techniques used in Clauset et al. (2008) or Faust (2008).

Acknowledgements

We thank the anonymous referees for their comments and suggestions in improving the presentation of this paper.

Supported in part by NSERC Discovery Grant RGPIN 327587-09.

Appendix A.

In order to discuss the proof of Theorem 3.3, we require the definitions of some related problems.

Problem 2. Cograph Deletion(G, k): Given a graph G and an integer k, is there a set S of at most k edges that can be deleted from G so that $G - S$ is P_4-free?

Problem 3. Cograph Editing(G, k): Given a graph G and an integer k, is there a set S of edge deletions and a set T of edge additions such that $|S| + |T| \leq k$ and $G - S + T$ is P_4-free?

Problem 4. Exact 3-Cover: Given a set of elements $S = \{s_1, s_2, \ldots, s_n\}$ and a collection $\mathcal{C} = \{S_1, S_2, S_3, \ldots, S_m\}$ of subsets of S with $|S_i| = 3$, is there a subcollection $\mathcal{T} \subseteq \mathcal{C}$ such that the union of all 3-sets in $\mathcal{T}$ is S, and every $s_i \in S$ is in a unique $S_j \in \mathcal{T}$?

We provide the details of a proof of Theorem 3.3 here by making some observations on existing related results for P_4 deletions.

Recall:

Problem 1. Quasi-Threshold Editing(G, k): Given a graph G and an integer k, is there a set S of edge deletions and a set T of edge additions such that $|S| + |T| \leq k$ and $G - S + T$ is (P_4, C_4)-free?

The polytime vs NP-completeness status of Quasi-Threshold Editing was left open by Burzyn et al. (2006) and also by Mancini (2008). Interestingly, it was also stated open in Liu et al. (2011) while their proof for NP-Completeness of Cograph Editing can be adapted to prove NP-completeness for Quasi-Threshold Editing using some extra observations.

Proof. (of Theorem 3.3) We use the same construction and notation as those in El-Mallah and Colbourn (1988), which was also used by Liu et al. (2011).

Let $S = \{s_1, s_2, \ldots, s_n\}$ and $\mathcal{C} = \{S_1, S_2, S_3, \ldots, S_m\}$ be an instance of Exact 3-Cover. Since each set $S_i \in \mathcal{C}$ contains three elements from S, an exact 3-cover of S would use exactly $n/3$ sets from $\mathcal{C}$. We let $n = 3t$ and $r = \binom{3t}{2}$. Using the same construction and notation from El-Mallah and Colbourn (1988) and Liu et al. (2011), we construct an instance of Quasi-Threshold Editing as follows.

- each s_i is a vertex, and the set S of these induces a clique,
- for every $S_j \in \mathcal{C}$, create two cliques X_j and Y_j such that $|X_j| = r$ and $|Y_j| = q$, where $q = 9(m - t)r + 3(r - 3t)$,
- each of the three elements s_a, s_b, s_c of S_j is adjacent to every $x \in X_j$ and every $x \in X_j$ is adjacent to every $y \in Y_j$,

- no other edges exist in this graph.

The parameter to this instance of Quasi-Threshold Editing is $k = q/3 = 3(m-t)r + (r-3t)$. This construction is depicted in Fig. 12.

We note that if the instance of exact 3-cover is nontrivial (if some s_i exists in at least two 3-sets) this constructed graph is not a quasi-threshold graph since there are many P_4s, for instance, starting in some Y_j, adjacent to a vertex in X_j, adjacent to s_i, and adjacent to another X_k. There are no induced C_4s in this graph, however.

First, we prove that if we have a solution to the Exact 3-Cover instance, we can find at most k edge edits to turn this constructed graph into a quasi-threshold graph. Say that $\mathcal{C}'$ is a collection of t subsets of $\mathcal{C}$ such that the union of subsets in $\mathcal{C}'$ is S. For every pair s_i and s_j in S, delete the edge joining s_i and s_j if they do not coexist in a 3-set S_i in the solution $\mathcal{C}'$. These amount to $r-3t$ edge deletions. Further, delete any edges from an X_i to S if S_i is not in $\mathcal{C}'$. This adds another $3(m-t)r$ deletions. In total, this gives $3(m-t)r + r - 3t = k$ edge edits, resulting in a (P_4, C_4)-free graph.

Note that a Quasi-Threshold Editing set of size at most k is also a cograph editing set of size at most k. Exactly the same argument used in Liu et al. (2011) can be used to show that the editing set contains edge deletions only. For the sake of completeness, we include the proof here.

Assume we have a Quasi-Threshold Editing set E' of size at most k and that the modified graph G' is (P_4, C_4)-free. Call an *affected* vertex to be a vertex with at least one incident edge that was modified by the k-edge edit set. Since each edge edit is incident on two vertices, there are at most $2k$ affected vertices.

Since $|Y_i| = 9(m-t)r + 3(r-3t) = 9(m-(n/3))\binom{n}{2} + 3\left(\binom{n}{2} - n\right) > 2\binom{n}{2} \geq 2k$, each Y_i set contains an unaffected vertex. We show that the edge edit set E' does not contain any edge additions.

Claim 1. *E' contains no edge from $X_i \cup Y_i$ to $X_j \cup Y_j$.*

Proof. Assume there is an edge $u = v_i v_j$ from $X_i \cup Y_i$ to $X_j \cup Y_j$, with $i \neq j$ as $X_i \cup Y_i$ is already a clique. Then let $y_i \in Y_i$ and $y_j \in Y_j$ be unaffected vertices. Since $v_i and v_j$ are affected by u, they are distinct from y_i and y_j. It is readily seen that $y_i v_i v_j y_j$ is a P_4, contradicting the fact that G' is quasi-threshold, so there can be no such edges. □

Claim 2. *Every vertex s_i in G' has at most one X_j whose vertices are adjacent to s_i.*

Proof. Assume s_i was adjacent to $x_p \in X_p$ and $x_q \in X_q$ where $p \neq q$. From the previous claim, x_p is not adjacent in x_q in G'. Let $y \in Y_p$ be an unaffected vertex. Then $y_p x_p s_i x_q$ is a P_4, contradicting the fact that G' is quasi-threshold. □

Claim 3. *If in G' we have that s_i is adjacent to X_p and s_j is adjacent to X_q, then E' must delete the edge $s_i s_j$.*

Proof. Since the previous claim shows that each s_i is adjacent to at most one X-set, we have that s_i is adjacent to X_p and so cannot be adjacent to X_q. Similarly, s_j cannot be adjacent to X_p. Since the first claim shows there is no edge from X_p to X_q, we have a P_4 from X_p to s_i to s_j to X_q, unless $s_i s_j$ is a deleted edge. □

Claim 4. *If E' is an optimal edge-edit set, then E' does not add any edge from Y_i to s_j.*

Proof. Assume there is some $y_i \in Y_i$ that is adjacent to s_j in the modified graph G'.

If s_j is adjacent to X_i, then the connected component containing s_j in the graph G' must be the vertex set $Y_i \cup X_i \cup \{s_j\}$. If there is an edge joining some y_i to s_j, then this edge can be removed from E', yielding a better edge-edit solution, since the graph induced by $Y_i \cup X_i \cup \{s_j\}$ in G is already (P_4, C_4)-free.

On the other hand, if s_j is adjacent to some X_p where $p \neq i$ then consider an unaffected vertex $y_p \in Y_p$. Using a vertex $x_p \in X_p$ which is adjacent to s_j as well as y_p, we find the P_4 $y_p x_p s_j y_i$.

Finally, if s_j is not adjacent to any X-set in G', then we either find a P_4 $s_j y_i x_i s_q$ in the case that X_i is adjacent to s_q. If X_i is not adjacent to any $s \in S$, then $\{s_j\} \cup Y_i \cup X_i$ is a connected component in G', and the added edge from s_j to y_i can be removed from E' yielding a better edit set. □

Claim 5. *If E' is an edge edit set of G such that the modified graph G' is quasi-threshold with $|E'| = k$, then E' either has no edge additions or else it can be improved to a smaller edge edit set E'', $|E''| < |E'|$ using only edge deletions.*

Proof. This follows from the previous set of claims, as we have shown that every possibility for an edge addition is either impossible or unnecessary. □

Claim 6. *If E' is an edge edit of set of G such that the modified graph G' is quasi-threshold with $|E'| = k$, then we can find a collection $\mathcal{C}'$ of t 3-sets which is an exact cover of S.*

Proof. Recall that $S = \{s_1, s_2, \ldots, s_n\}$ and $\mathcal{C} = \{S_1, S_2, S_3, \ldots, S_m\}$. Since each s_i is adjacent to at most one X_j at most $t = n/3$ groups of 3 s_i-vertices can remain adjacent to X_j sets, and so $(m-t)$ of the X_j sets must be disconnected from S. Since each X_j is adjacent to 3 vertices in S, and $|X_j| = r = \binom{n}{2}$, at least $3(m-t)r$ edges must be deleted from $\bigcup X_i$ to S by Claim 2.

Furthermore, Claim 3 implies that there are deletions within the S set. The most number of edges that can be left in S will partition S into $t = n/3$ triangles (so these triangles contain $3t$ edges). Since S originally has $\binom{n}{2} = r$ edges, Claim 3 implies at least $r - 3t$ edges were deleted.

The remainder of the claims show that there are no other edge edits necessary, and so the number of edge edits $|E'|$ is at least $3(m-t)r + r - 3t$ edges. But E' was taken to be a deletion set of at most size $k = 3(m-t)r + (r-3t)$, so it must be that $|E'| = 3(m-t)r + (r-3t)$. It follows that the modified graph G' has exactly t triangles in S and thus gives the required exact cover of S. □

This completes the proof of the theorem. □

References

Alba, R.D., 1973. A graph-theoretic definition of a sociometric clique. Journal of Mathematical Sociology 3, 113–126.

Balasundaram, B., Butenko, S., Hicks, I.V., Sachdeva, S., 2011. Clique relaxations in social network analysis: the maximum k-plex problem. Operations Research 59 (1), 133–142.

Balasundaram, B., Butenko, S., Trukhanov, S., 2005. Novel approaches for analyzing biological networks. Journal of Combinatorial Optimization 10, 23–39.

Bansal, N., Blum, A., Chawla, S., 2004. Correlation clustering. Machine Learning 56, 89–113.

Böcker, S., Briesemeister, S., Klau, G.W., 2011. Exact algorithms for cluster editing: evaluation and experiments. Algorithmica 60 (2), 316–334.

Brandstädt, A., Le, V.B., Spinrad, J.P., 1999. Graph Classes: A Survey. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Burzyn, P., Bonomo, F., Durán, G., 2006. NP-completeness results for edge modification problems. Discrete Applied Mathematics 154 (13), 1824–1844.

Cai, L., 1996. Fixed-parameter tractability of graph modification problems for hereditary properties. Information Processing Letters 58 (4), 171–176.

Chu, F.P.M., June 2008. A simple linear time certifying LBFS-based algorithm for recognizing trivially perfect graphs and their complements. Information Processing Letters 107, 7–12.

Chvátal, V., Hammer, P.L., 1977. Aggregation of inequalities in integer programming. In: Hammer, P.L., Johnson, E.L., G.L.Nemhauser, B.H.K. (Eds.), Studies in Integer Programming. Vol. 1 of Annals of Discrete Mathematics. Elsevier, pp. 145–162.

Clauset, A., Moore, C., Newman, M.E.J., 2008. Hierarchical structure and the prediction of missing links in networks. Nature 453, 98–101.

Cozzens, M.B., Kelleher, L.L., 1990. Dominating cliques in graphs. Discrete Mathematics 86 (1–3), 101–116.

Davis, J.A., Leinhardt, S., 1967. The structure of positive interpersonal relations in small groups. In: Berger Jr., J., B.Anderson, M.Z. (Eds.), In: EBST Sociological Theories in Progress, 2. Houghton Mifflin, Boston, pp. 218–251.
Dawah, H.A., Hawkins, B.A., Claridge, M.F., 1995. Structure of parasitoid communities of grass-feeding chalcid wasps. Journal of Animal Ecology 64, 708–720.
Donnelly, S., Isaak, G., 1999. Hamiltonian powers in threshold and arborescent comparability graphs. Discrete Mathematics 202 (1–3), 33–44.
El-Mallah, E.S., Colbourn, C.J., 1988. Edge deletion problems: properties defined by weakly connected forbidden subgraphs. In: Proc. Eighteenth Southeastern Conference on Combinatorics, Graph Theory, and Computing Congressus Numerantium, vol. 61, pp. 275–285.
Evans, T.S., 2010. Clique graphs and overlapping communities. Journal of Statistical Mechanics: Theory and Experiment 12, P12037.
Falzon, L., 2000. Determining groups from the clique structure in large social networks. Social Networks 22, 159–172.
Faust, K., 2008. Triadic configurations in limited choice sociometric networks: empirical and theoretical results. Social Networks 30 (4), 273–282.
Fellows, M.R., Guo, J., Komusiewicz, C., Niedermeier, R., Uhlmann, J., 2011. Graph-based data clustering with overlaps. Discrete Optimization 8 (1), 2–17.
Fortunato, S., 2010. Community detection in graphs. Physics Reports 486 (3–5), 75–174.
Freeman, L.C., 1977. A set of measures of centrality based upon betweenness. Sociometry 40, 35–41.
Freeman, L.C., 1978. Centrality in social networks: conceptual clarification. Social Networks 1, 215–239.
Freeman, L.C., 1996. Cliques, galois lattices, and the structure of human social groups. Social Networks 18, 173–187.
Girvan, M., Newman, M.E.J., 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of United States of America 99 (12), 7821–7826.
Golumbic, M.C., 1978. Trivially perfect graphs. Discrete Mathematics 24 (1), 105–107.
Golumbic, M.C., 2004. Foreword 2004: the annals edition. In: Golumbic, M.C. (Ed.), Algorithmic Graph Theory and Perfect Graphs. Vol. 57 of Annals of Discrete Mathematics. Elsevier, pp. xiii–xiv.
Granovetter, M.S., 1973. The strength of weak ties. American Journal of Sociology 78 (6), 1360–1380.
Guo, J., 2007. Problem kernels for NP-complete edge deletion problems: split and related graphs. In: ISAAC, pp. 915–926.
Hastings, M.B., Sep 2006. Community detection as an inference problem. Physical Reviews E. 74 (3), 035102, http://link.aps.org/doi/10.1103/PhysRevE.74.035102.
Homans, G.C., 1950. The Human Group. Harcourt, Brace and Co.
Karrer, B., Levina, E., Newman, M.E.J., Apr 2008. Robustness of community structure in networks. Physical Reviews E 77 (4), 046119.
Kazienko, P., Musial, K., 2007. Assessment of personal importance based on social networks. In: Gelbukh, A., Kuri Morales, N. (Eds.), MICAI 2007: Advances in Artificial Intelligence. Vol. 4827 of Lecture Notes in Computer Science. Springer, Berlin/Heidelberg, pp. 529–539.
Knuth, D.E., 1993. The Stanford GraphBase: A Platform for Combinatorial Computing. Addison-Wesley, Reading, MA.
Lemmouchi, S., Haddad, M., Kheddouci, H., 2009. Robustness of emerged community in social network. In: Proceedings of the International Conference on Management of Emergent Digital EcoSystems. MEDES'09, pp. 78:477–78:479.
Liu, Y., Wang, J., Guo, J., Chen, J., 2012. Complexity and parameterized algorithms for cograph editing. Theoretical Computer Science 461, 45–54.
Luccio, F., Sami, M., 1969. On the decomposition of networks in minimally interconnected subnetworks. IEEE Transactions on Circuit Theory 16, 184–188.
Luce, R.D., 1950. Connectivity and generalized cliques in sociometric group structure. Psychometrika 15, 169–190.
Lusseau, D., 2003. The emergent properties of a dolphin social network. Proceedings of the Royal Society of London Series B: Biological Sciences 270, S186–S218.
Mancini, F., 2008. Graph modification problems related to graph classes. University of Bergen (PhD thesis).
Mishra, N., Schreiber, R., Stanton, I., Tarjan, R.E., 2008. Finding strongly knit clusters in social networks. Internet Mathematics 5 (1–2), 155–174.
Mokken, R.J., 1979. Cliques, clubs and clans. Quality and Quantity 13, 161–173.
Nastos, J., Gao, Y.,2010. A novel branching strategy for parameterized graph modification problems. In: Proceedings of the 4th International Conference on Combinatorial Optimization and Applications – Volume Part II. COCOA'10. Springer-Verlag, pp. 332–346.
Nastos, J., Gao, Y., 2012. Bounded search tree algorithms for parameterized cograph deletion: efficient branching rules by exploiting structures of special graph classes. Discrete Mathematics, Algorithms and Applications 4 (1).
Newman, M.E.J., 2010. Networks: An Introduction. Oxford University Press.
Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Physical Review E 69.
Palla, G., Derényi, I., Farkas, I., Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814.
Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D., 2004. Defining and identifying communities in networks. Proceedings of the National Academy of Sciences of United States of America 101, 2658–2663.
Ravasz, E., Barabási, A.-L., Feb 2003. Hierarchical organization in complex networks. Physical Reviews E 67, 026112.
Reed, B.A., 1987. A semi-strong perfect graph theorem. Journal of Combinatorial Theory, Series B 43 (2), 223–240.
Schaeffer, S.E., 2007. Graph clustering. Computer Science Review 1, 27–64.
Seidman, S.B., Foster, B.L., 1978. A graph theoretic generalization of the clique concept. Journal of Mathematical Sociology 6, 139–154.
Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58 (301), 236–244.
White, S., Smyth, P., 2005. A spectral clustering approach to finding communities in graphs. Proceedings of the Fifth SIAM International Conference on Data Mining 119, 76–78.
Wolk, E.S., 1962. The comparability graph of a tree. Proceedings of the American Mathematical Society 13, 789–795.
Yan, J.-H., Chen, J.-J., Chang, G.J., 1996. Quasi-threshold graphs. Discrete Applied Mathematics 69 (3), 247–255.
Zachary, W.W., 1977. An information flow model for conflict and fission in small groups. Journal of Anthropological Research 33, 452–473.