Line graphs, link partitions, and overlapping communities

Evans, T.S.; Lambiotte, R.

Published in:
Physical Review E - Statistical, Nonlinear, and Soft Matter Physics

DOI:
10.1103/PhysRevE.80.016105

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

Line graphs, link partitions, and overlapping communities

T. S. Evans[1,2] and R. Lambiotte[1]
[1]*Institute for Mathematical Sciences, Imperial College London, SW7 2PG London, United Kingdom*
[2]*Theoretical Physics, Imperial College London, SW7 2AZ London, United Kingdom*
(Received 12 March 2009; published 9 July 2009)

In this paper, we use a partition of the links of a network in order to uncover its community structure. This approach allows for communities to overlap at nodes so that nodes may be in more than one community. We do this by making a node partition of the line graph of the original network. In this way we show that any algorithm that produces a partition of nodes can be used to produce a partition of links. We discuss the role of the degree heterogeneity and propose a weighted version of the line graph in order to account for this.

DOI: 10.1103/PhysRevE.80.016105 PACS number(s): 89.75.−k, 02.50.Le, 05.50.+q, 75.10.Hk

I. INTRODUCTION

Finding hidden patterns or regularities in data sets is a universal problem that has a long tradition in many disciplines from computer science [1] to social sciences [2]. For example, when the data set can be represented as a graph, i.e., a set of elements and their pairwise relationships, one often searches for tightly knit sets of nodes usually called communities or modules. The identification of such communities is particularly crucial for large network data sets that require new mathematical tools and computer algorithms for their interpretation. Most community detection methods find a partition of the set of nodes where most of the links are concentrated within the communities [3,4]. Here the communities are the elements of the partition and so each node is in one and only one community.

A popular class of algorithms seeks to optimize the modularity Q of the partition of the nodes of a graph G [5–9]. The simplest definition of modularity for an undirected graph, i.e., the adjacency matrix $\mathbf{A}$ is symmetric, is [10]

$$Q(\mathbf{A}) = \frac{1}{W}\sum_{C\in\mathcal{P}}\sum_{i,j\in C}\left[A_{ij} - \frac{k_i k_j}{W}\right], \tag{1}$$

where $W=\Sigma_{i,j}A_{ij}$ and $k_i=\Sigma_j A_{ij}$ is the degree of node i. The indices i and j run over the N nodes of the graph G. The index C runs over the communities of the partition $\mathcal{P}$. Modularity counts the number of links between all pairs of nodes belonging to the same community and compares it to the expected number of such links for an equivalent random graph in which the degree of all nodes has been left unchanged. By construction $|Q|\leq 1$ with larger Q indicating that more links remain within communities then would be expected in the random model. Uncovering a node partition that optimizes modularity is therefore likely to produce useful communities.

This node partitioning approach has, however, the drawback that nodes are attributed to only one community, which may be an undesirable constraint for networks made of highly overlapping communities. This would be the case, for instance, for social networks, where individuals typically belong to different communities, each characterized by a certain type of relation, e.g., friendship, family, or work. In scientific collaboration networks (for example [11]), authors may belong to different research groups characterized by different research interests. Such intercommunity individuals are often of great interest as they broker the flow of information between otherwise disconnected contacts, thereby connecting people with different ideas, interests, and perspectives [12,13].

Only a few alternative approaches have been proposed in order to uncover overlapping communities of nodes, for example [14–16]. Our suggestion is to define communities as a partition of the links rather than of the set of nodes. A node may then have links belonging to several communities and in this it belongs to several communities. The central node in a bow tie graph is a simple example; see Fig. 1. This link partition approach should be especially efficient in situations when the nodes of a network are connected by different types of links, i.e., in situations where the nodes are heterogeneous while the links are very homogeneous. In the case of the social network mentioned above, this would occur when the friendship network and work network of individuals only have a very small overlap.

This paper is organized as follows. In Sec. II, we review a definition of modularity that uses the statistical properties of a dynamical process taking place on the nodes of a graph. In Sec. III, we propose three dynamical processes taking place on the links of the graph and derive their corresponding modularities, now defined for a partition of the links of a network. To do so, we make connections to the concept of a line graph and with the projection of bipartite networks. In Sec. IV, we optimize the three modularities for some examples and interpret our results. In Sec. V we conclude and propose ways to improve our method.

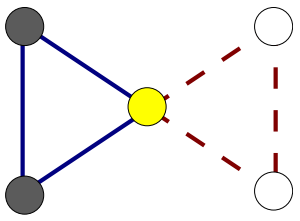

FIG. 1. (Color online) By partitioning the links of a network into communities, one may uncover overlapping communities for the nodes by noting that a node belongs to the communities of its links. In this toy example, a meaningful partition consists in dividing the links into two groups (straight blue lines and the dashed red lines). In that case, the central node belongs to the two communities because it is at the interface between these link communities.

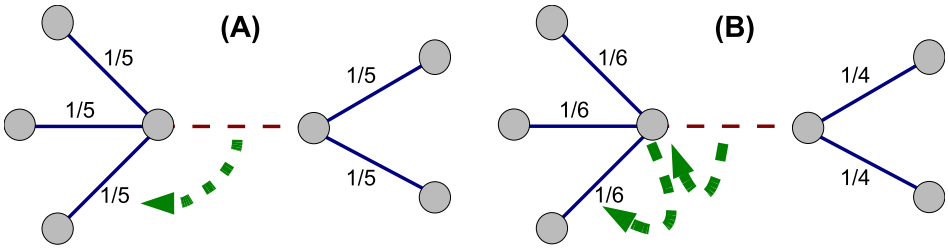

FIG. 2. (Color online) Illustration of the two types of random walk considered in this paper. In both cases, the walkers are situated on the links of a graph, here starting from the central red dashed link. In (a) the link-link random walk is shown where the walker jumps (the green dashed arrows) to any of the adjacent links with equal probability. In (b) a link-node-link random walk is illustrated. In this case the walker moves first to a neighboring node with equal probability and then moves on to a new link chosen with equal probability from those new links incident at the node.

II. DYNAMICAL FORMULATION OF MODULARITY

To motivate our link partition quality function, let us first consider how to interpret the usual modularity Q [Eq. (1)] in terms of a random walker moving on the nodes [17,18]. Suppose that the density of random walkers on node i at step n is $p_{i;n}$ and the dynamics is given by

$$p_{i;n+1} = \sum_j \frac{A_{ij}}{k_j} p_{j;n}. \tag{2}$$

From now on, we will only consider networks that are undirected (the adjacency matrix is symmetric), connected (there exists a path between all pairs of nodes), nonbipartite (it is not possible to divide the network into two sets of nodes such that there is no link between nodes of the same set), and simple (without self-loops nor multiple links). If the first three conditions are respected, it is easy to show [19] that the stationary solution of the dynamics is generically given by $p_i^* = k_i/W$.

Let us now consider a node partition $\mathcal{P}$ of the network and focus on one community $C \in \mathcal{P}$. If the system is at equilibrium, it is straightforward to show that the probability a random walker is in C on two successive time steps is

$$\sum_{i,j \in C} \frac{A_{ij}}{k_j} \frac{k_j}{W}, \tag{3}$$

while the probability of finding two independent walkers at nodes in C are

$$\sum_{i,j \in C} \frac{k_i k_j}{(W)^2}. \tag{4}$$

This observation allows us to reinterpret Q as a summation over the communities of the difference of these two probabilities. This interpretation suggests natural generalizations of modularity allowing to tune its resolution. Indeed, Q is based on paths of length one but it can readily be generalized to paths of arbitrary length as

$$R(\mathbf{A},n) = \frac{1}{W} \sum_{C \in \mathcal{P}} \sum_{i,j \in C} \left[(T^n)_{ij} k_j - \frac{k_i k_j}{W} \right], \tag{5}$$

where $T_{ij} = A_{ij}/k_j$. This quantity is called the stability of the partition [17]. Because k_j is an eigenvector of eigenvalue one of $\mathbf{T}$, one can show that the symmetric matrix $X(n)_{ij} = (T^n)_{ij} k_j$ corresponds to a time-dependent graph where the degree of node i is always equal to k_i. Therefore $R(\mathbf{A},n)$ can be interpreted as the modularity of $X(n)_{ij}$, a matrix that connects more and more distant nodes of the original adjacency matrix A as time n grows [18]. It can be shown that optimizing Eq. (5) typically leads to partitions made of larger and larger communities for increasing times and that the optimal partition when $n \to \infty$ is made of two communities [17,18].

III. LINK PARTITION

A. Random walking the links

The above discussion suggests that we should look at a random walker moving on the links of network in order to define the quality of a link partition. Such a walker would therefore be located on the links instead of the nodes at each time n and move between adjacent links, i.e., links having one node in common. In the case of the random walk on the nodes [Eq. (2)], a walker at node i follows one of its links with probability $1/k_i$, i.e., all links are treated equally. However, a link between nodes i and j is characterized by two quantities k_i and k_j, so a random walk on the links is more subtle. In the following, we will focus on two different types of dynamical processes that account differently for the degrees k_i and k_j (see Fig. 2).

In the first process, a walker jumps with the same probability $1/(k_i + k_j - 2)$ to one of the links leaving i and j. When $k_i \neq k_j$, the walker goes with a different probability through i or j, and we therefore call this process a "link-link random walk" [see Fig. 2(a)].

In the second process, a walker jumps to one of the two nodes to which it is attached, say i, then moves to a link attached to that node (excluding the link it came from). Thus it will arrive at a link leaving node i with a probability $1/[2(k_i - 1)]$, and similarly it will arrive at a link attached to the other node j with probability $1/[2(k_j - 1)]$. We will refer to this as a "link-node-link random walk" [see Fig. 2(b)]. This process is well defined unless the link is a leaf, namely, one of its extremities has a degree 1, say i. In that case, the walker will jump with a probability $1/(k_j - 1)$ to one of the links leaving j.

These two types of dynamics are different in general except if the degrees at the extremities i and j of each link are

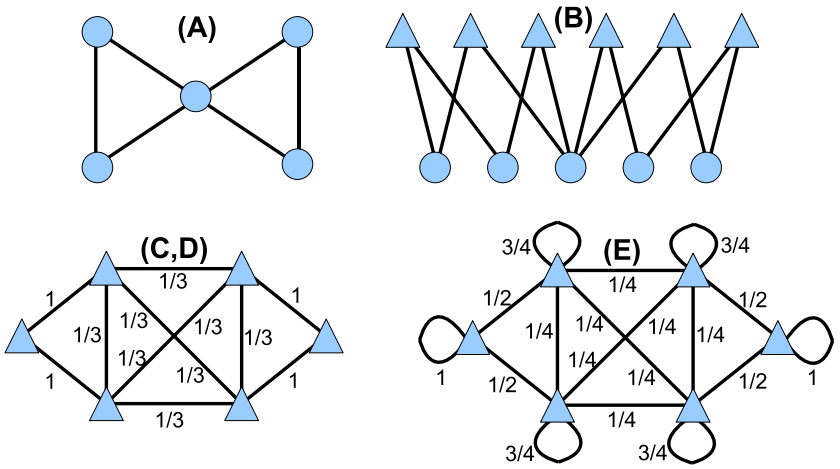

FIG. 3. (Color online) The information of the bow tie graph in (a), as encoded by the adjacency matrix **A** of Eq. (7), has other equivalent graph representations. In (b) the incidence matrix [**B** of Eq. (7)] of the bow tie is shown as a bipartite network, the incidence graph $I(G)$. The line graph of the bow tie, $L(G)$, is the unweighted version of the graph labeled (c),(d) with adjacency matrix **C** of Eq. (8). The weighted version in diagram (c),(d) has an adjacency matrix **D** of Eq. (11). The weighted line graph with self-loops labeled (e) has an adjacency matrix **E** of Eq. (14). Circles represent entities that correspond to nodes of the original graph, while triangles come from links in the original graph.

equal. In the case of a connected graph, this condition is equivalent to demanding that the graph is regular, i.e., the degree of all the nodes is a constant. When this condition is not respected, the link-link random walk favors the passage of the walker through the extremity having the largest degree. The difference between the two processes will be maximal when the network is strongly disassortative, namely, when links typically relate nodes with very different degrees [20].

B. Projecting the incidence matrix

1. Bipartite structure

In order to study these two types of random walk more carefully, it is useful to represent a network G by its incidence matrix **B**. The elements $B_{i\alpha}$ of this $N\times L$ matrix (L is the number of links) are equal to 1 if link α is related to node i and 0 otherwise. The incidence matrix of G may be seen as the adjacency matrix of a bipartite network $I(G)$ [see Fig. 3(b)], the incidence graph[1] of G where the two types of nodes correspond to the nodes and the links of the original graph G. By construction, all the information of the graph is incorporated in **B**. For instance, the degree k_i of a node i and the number of nodes k_α attached to a link α (always equal to 2) are given by

$$k_i = \sum_\alpha B_{i\alpha}, \quad k_\alpha = \sum_i B_{i\alpha}. \tag{6}$$

The $N\times N$ adjacency matrix **A** of the graph G can also be obtained

$$A_{ij} = \sum_\alpha B_{i\alpha}B_{j\alpha} - k_i\delta_{ij}. \tag{7}$$

This operation (7) can be interpreted as a projection of the bipartite incidence graph $I(G)$ onto the unipartite network G [21,22]. In a similar way, an adjacency matrix for the links can be obtained by projecting the bipartite network onto its links. In the following, we will focus on two standard types of projection that, as we will show, are directly related to the two random walks introduced above.

2. Line graph

The simplest way to project a bipartite graph consists of taking all the nodes of one type for the nodes of the projected graph. A link is added between two nodes in this projected graph if these two nodes had at least one node of the other type in common in the original bipartite graph. Operation (7) is of this type. When applied to the links α of the graph G, the second type of vertex in the bipartite incidence graph $I(G)$, it leads to the $L\times L$ adjacency matrix C whose elements are

$$C_{\alpha\beta} = \sum_i B_{i\alpha}B_{i\beta}(1-\delta_{\alpha\beta}). \tag{8}$$

It is easy to verify that this adjacency matrix is symmetric and that its elements are equal to 1 if two links have one node in common, and zero otherwise. It is interesting to note that this adjacency matrix corresponds to another well-known graph, usually called the *line graph* of G [23] and denoted by $L(G)$ [see Fig. 3(c)]. It is a simple graph with L nodes. By construction, each node i of degree k_i of the original graph G corresponds to a k_i fully connected clique in $L(G)$. Thus it has $\Sigma_i k_i(k_i-1)/2 = O(\langle k^2\rangle N)$ links. Line graphs have been studied extensively and among their well-known properties, Whitney's uniqueness theorem states that the structure of G can be recovered completely from its line

[1] An incidence graph is usually defined in terms of the incidence of a set of lines with a set of points in a Euclidean space of finite dimension. Here we have a special case where we embed our graph G in some Euclidean space of no particular interest and each link of G is a line that always intersects with exactly two points.

graph $L(G)$ for any graph other than a triangle or a star network of four nodes [24]. This result implies that projecting the incidence matrix onto $L(G)$ does not lead to any loss of information from the network structure. This is a remarkable result that is not generally true when projecting generic bipartite networks.

It is now straightforward to express the dynamics of link-link random walk [Fig. 2(a)] in terms of the projected adjacency matrix C,

$$p_{\alpha;n+1} = \sum_{\beta} \frac{C_{\alpha\beta}}{k_\beta} p_{\beta;n}. \tag{9}$$

Now $p_{\alpha;n}$ is the density of random walkers on link α at step n, $k_\alpha = \Sigma_\beta C_{\alpha\beta} = (k_i + k_j - 2)$ and where i and j are the extremities of α. This dynamical process therefore only depends on the sum of the degrees i and j. The stationary solution is found to be $p^*_\alpha = k_\alpha / W$, where $W = \Sigma_{\alpha\beta} C_{\alpha\beta}$. When G is simple, then $W = \Sigma_i (k_i - 1) k_i$. By reapplying the steps described in [18], it is now straightforward to derive a quality function for the link partition $\mathcal{P}$ of the graph G,

$$Q(\mathbf{C}) = \frac{1}{W} \sum_{C \in \mathcal{P}} \sum_{\alpha,\beta \in C} \left[C_{\alpha\beta} - \frac{k_\alpha k_\beta}{W} \right]. \tag{10}$$

This is just the usual modularity (1) for a graph with adjacency matrix $\mathbf{C}$.

As we noted, a single node i in G leads to a connected clique of $k_i(k_i-1)/2$ links in the line graph $L(G)$. This seems to suggest that the line graph $L(G)$ gives too much prominence to the high degree nodes of the original graph G. Our response is to define a weighted line graph whose links are scaled by a factor of $O(1/k_i)$.

3. Weighted line graph

In order to derive the quality of a link partition associated to the link-node-link random walk, it is useful to project the incidence matrix in a different way and to define another graph $D(G)$ with a symmetric adjacency matrix given by

$$D_{\alpha\beta} = \sum_{i,k_i>1} \frac{B_{i\alpha} B_{i\beta}}{k_i - 1} (1 - \delta_{\alpha\beta}). \tag{11}$$

This weighted line graph has the intuitive property that the degree $k_\alpha = \Sigma_\beta D_{\alpha\beta}$ of a link α is equal to 2 (a link always has two extremities) unless α is a leaf in G (then $k_\alpha = 1$ except for one trivial case). For example this weighted line graph of the bow tie network is shown in Fig. 3(d). Only if G is regular will this weighted line graph $D(G)$ be equivalent (up to an overall scale) to the original unweighted line graph $L(G)$.

This construction is a well-known method for projecting bipartite networks. For instance in the case of collaboration networks [11] the $(k_i - 1)$ normalization is justified by the desire that two authors should be less connected if they wrote a joint paper with many co-authors than a paper with few authors.

This weighted line graph allows us to write the dynamics of the link-node-link random walk in a natural way

$$p_{\alpha;n+1} = \sum_{\beta} \frac{D_{\alpha\beta}}{k_\beta} p_{\beta;n} \tag{12}$$

and, by reusing the above arguments to define another quality function for the link partition $\mathcal{P}$ of a graph

$$Q(\mathbf{D}) = \frac{1}{W} \sum_{C \in \mathcal{P}} \sum_{\alpha,\beta \in C} \left[D_{\alpha\beta} - \frac{k_\alpha k_\beta}{W} \right], \tag{13}$$

where $W = \Sigma_{\alpha\beta} D_{\alpha\beta} = 2L - L_{\text{leaf}}$ is twice the number of links L minus the number of leaves in the original graph G, L_{leaf}. Again, this is the same functional form as the usual modularity, $Q(\mathbf{A})$ of Eq. (1), only the adjacency matrix has changed.

C. Projection of a node random walk

The random walks proposed in the previous sections have been defined on the line graph and therefore consist of walkers moving among adjacent links of the original graph G. However, such processes cannot be related to the original random walk (3) on the nodes of G, because a walker moving on links can pass at two subsequent steps through the same node of G while such self-loops are forbidden in Eq. (3). This observation suggests an alternative approach where the dynamics would be driven by the original random walk (3) but would be projected on the links of the network. To do so, let us assume that a walker has not moved yet and is located at node i. In that case, it is reasonable to assume that all the neighboring links of i are connected by a weight $1/k_i$. The corresponding adjacency matrix E for the links is therefore given by

$$E_{\alpha\beta} = \sum_{i,k_i>0} \frac{B_{i\alpha} B_{i\beta}}{k_i} \tag{14}$$

and is based on an unconstrained unbiased two-step random walk on the bipartite incidence graph $I(G)$ [2]. Unlike our previous line graph constructions, $\mathbf{C}$ of Eq. (8) and $\mathbf{D}$ of Eq. (11), this weighted line graph $E(G)$ has self-loops. It is illustrated for the bow tie graph in Fig. 3(e). All nodes α in $E(G)$ have strength 2, $\Sigma_\beta E_{\alpha\beta} = 2$, reflecting the fact that the links in the original graph G all have two ends.

$\mathbf{E}$ is constructed when a walker is located on a node and has not moved yet. The motion of the walker according to Eq. (3) generates a new adjacency matrix, $\mathbf{E}_1$, defined as

[2]One might also try to argue that since an undirected link is both incoming and outgoing, we might deem it appropriate to allow α to α transitions in the link-link walk of Fig. 2(a). That is, we could define an unweighted line graph with self-loops with adjacency matrix $\tilde{C}_{\alpha\beta} = \Sigma_i B_{i\alpha} B_{i\beta}$. Since it differs from the standard unweighted line graph $L(G)$ only by the addition of a self-loop to every node α, this can be interpreted within the scheme of [29] who add self-loops to control the number and size of communities found.

$$E_{1;\alpha\beta} = \sum_{i,k_i>0} \frac{B_{i\alpha}A_{ij}B_{i\beta}}{k_i k_j}, \tag{15}$$

where we note that $\mathbf{E}_1 = \mathbf{EE} - \mathbf{E}$. The corresponding graph is still regular with $k_\alpha = \Sigma_\beta E_{1;\alpha\beta} = 2$, and it is again weighted with self-loops. The quality function associated with this dynamics is simply

$$Q(E_1) = \frac{1}{W} \sum_{C\in\mathcal{P}} \sum_{\alpha,\beta\in C} \left[E_{1;\alpha\beta} - \frac{4}{W} \right], \tag{16}$$

where again $W = 2L$.

This quality function is particularly interesting because it has a simple relationship to the modularity of the original graph, $Q(\mathbf{A})$ of Eq. (1). To show this let us assign a weight $V_{\alpha c}$ representing the strength of the membership of link α in community c. Such weights may be defined and constrained in many ways. For instance, in a link partition we have $V_{\alpha c}V_{\alpha d} = \delta_{cd}$ for any α, i.e., every link α belongs to just one community. In order to translate $V_{\alpha c}$ into a community structure on the nodes, it is natural to use the incidence matrix $\mathbf{B}$ of Eq. (7) and to define the rectangular matrix V_{ic} through

$$V_{ic} = \sum_\alpha \frac{B_{i\alpha}}{k_i} V_{\alpha c}. \tag{17}$$

If $V_{\alpha c}$ is a link partition then the projected node community structure V_{ic} is simply the fraction of links in community c incident at node i. Also if $\Sigma_c V_{\alpha c} = 1$ then so is $\Sigma_c V_{ic} = 1$.

Now using the definition of the adjacency matrix in Eq. (7), we find that the modularity of the original graph G for some node community V_{ic} is

$$Q(E_1;\{V_{\alpha c}\}) = \frac{1}{W} \sum_{c,d} \sum_{\alpha,\beta} V_{\alpha c} \left[E_{1;\alpha\beta} - \frac{4}{W} \right] V_{\beta d} \tag{18}$$

$$= \frac{1}{W} \sum_{c,d} \sum_{i,j} V_{ic} \left[A_{ij} - \frac{k_i k_j}{W} \right] V_{jd} \tag{19}$$

$$= Q(\mathbf{A};\{V_{ic}\}). \tag{20}$$

Thus finding modularity optimal link partitions of the line graph with adjacency matrix $\mathbf{E}_1$ of Eq. (15) is equivalent to the optimization of the modularity of the original graph but with a different constraint on the node community V_{ic} from that imposed when finding node partitions.

IV. EMPIRICAL ANALYSIS

A. Methodology

In the previous sections, we have proposed three quality functions $Q(\mathbf{C})$, $Q(\mathbf{D})$, and $Q(\mathbf{E}_1)$ for the partition of the links of a network G. Each represents a different dynamical process and therefore explores the structure of the original graph G in a different way. In order to tune the resolution of the optimal partitions, it is straightforward to define the stabilities $R(\mathbf{C},n)$, $R(\mathbf{D},n)$, and $R(\mathbf{E}_1,n)$ of the three processes by generalizing the concept of modularity to paths of arbitrary length (see Sec. II). The optimal partitions of these quality functions can be found by applying standard modularity optimization algorithms to the corresponding line graphs. In this paper, we have used two different algorithms [7,8] and have verified that both algorithms give consistent results.

As a first check, let us look at the bow tie graph of Fig. 1. The optimization of the three quality functions $Q(\mathbf{C})$, $Q(\mathbf{D})$, and $Q(\mathbf{E}_1)$ lead to the expected partition into two triangles, with the values $Q(\mathbf{C}) = 0.1$, $Q(\mathbf{D}) = 0.278$, and $Q(\mathbf{E}_1) = 0.167$. In this case, the central node belongs equally to the two link communities, a situation that is a far superior way to split the network than a node partition. The best node partition gives $Q(\mathbf{A}) = 0.111$ when three nodes in one triangle form one community and the remaining two nodes form a second community.

In order to compare node partitions and link partitions in the following, we will use the idea of a "boundary link" and a "boundary node." A boundary link of a node partition is one that connects two nodes from different communities. We will then define a boundary node of a link partition to be a node that is connected to links from more than one link community. Thus the central node of the bow tie graph is a boundary node.

B. Karate club

A less contrived graph is the Karate club of Zachary [2], which is made of 34 members. Historically this split into two distinct factions. It is standard to compare the partition produced by a community detection method to the actual split of the club. The node partition having the largest value of modularity $Q(\mathbf{A}) = 0.420$ contains four communities, but the resolution can be lowered by optimizing the stability $R(\mathbf{A},n)$ for larger values of n. When n is large enough, the optimal partition is always made of two communities (see Fig. 4), e.g., $R(\mathbf{A},11) = 0.078$, which agree with Zachary's partition into "sink" and "source" communities [2] using the Ford-Fulkerson binary community algorithm [25].

The link partitions found by optimizing $Q(\mathbf{C}) = 0.5$, $Q(\mathbf{D}) = 0.53$, and $Q(\mathbf{E}_1) = 0.36$ are shown in Fig. 5. They are, respectively, made of four, seven, and three communities. These three partitions are consistent with the historical two-way split of the network, as the boundary links of the two-way partition of Fig. 4 are always connected to a boundary node of a link partition. In general, however, the three optimal partitions are as different as their corresponding dynamical processes are. The most striking difference is observed around node 1. In the node partition optimizing $Q(\mathbf{A})$, this node is connected to several boundary links and connects the community of nodes (5,6,7,11,17) to the rest of the network. Such a position is consistent with the link partitions obtained from $Q(\mathbf{D})$ and $Q(\mathbf{E}_1)$, but not with the link partition optimizing $Q(\mathbf{C})$. In this latter case, one observes that node 1 is rather the focus of one of the link communities on the left-hand side in Fig. 5. This difference originates from the high degree of node 1, which implies that a link-link random walk is biased to pass through this node (see Fig. 2) and therefore heavily connects its adjacent links. This is a general problem of the unweighted line graph $\mathbf{C}$ that gives too much emphasis to high degree nodes (also noted in [27]) and therefore tends

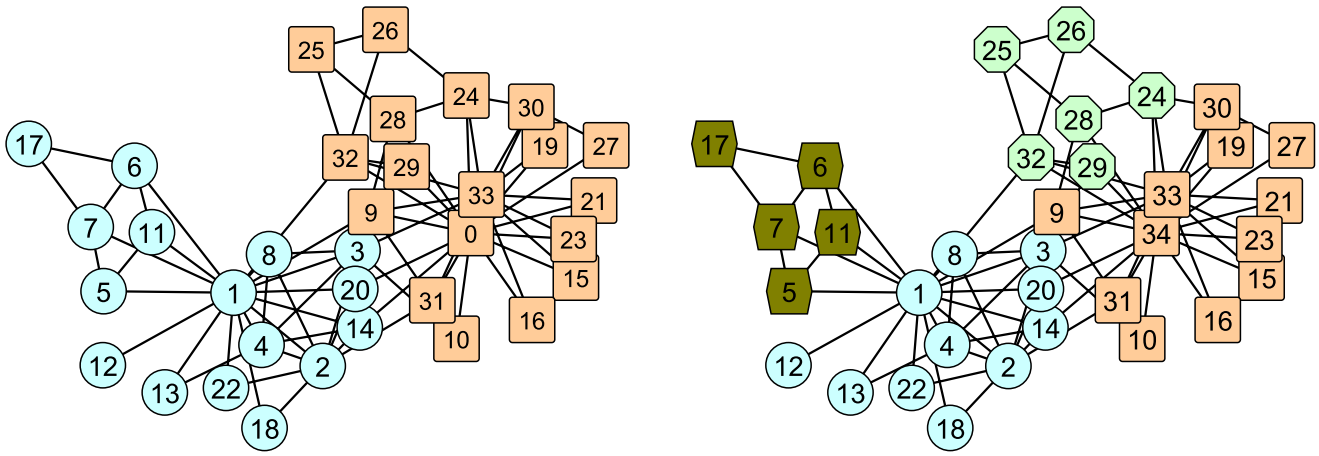

FIG. 4. (Color online) Optimal node partitions for the unweighted Karate club data of Zachary, notation as in [2]. On the left is the partition into two communities made by Zachary [2] using the Ford-Fulkerson binary community algorithm [25]. It is also produced by optimizing $R(\mathbf{A},11)$ of Eq. (5). The right-hand figure shows the node partition with optimal $Q(\mathbf{A})=0.420$ [26], which contains four communities.

to produces communities centered around hubs. Such a problem does not take place for the weighted line graphs $\mathbf{D}$ and $\mathbf{E}_1$, and in both these cases node 1 is a boundary node, part of several communities. The main difference between the optimal partitions of $Q(\mathbf{D})$ and $Q(\mathbf{E}_1)$ is the number of the communities in each, as expected because the line graph $\mathbf{E}_1$ connects more distance links of the original graph than $\mathbf{D}$. Let us also note that the optimal partition of $Q(\mathbf{E}_1)$ resembles very much the one of $Q(\mathbf{A})$, as suggested by Eq. (20).

Before concluding, let illustrate how longer random walks can be used to tune the resolution of the link partition. We focus on the weighted line graph $\mathbf{D}$, whose optimal partition into seven communities is difficult to compare against the standard two and four community node partitions of Fig. 4. Let us therefore focus on the stability $R(\mathbf{D},n)$, which is based on paths of length n of a random walker on $\mathbf{D}$. As expected, larger and larger communities are uncovered when n is increased and, when n is large enough, we obtain a two way link partition (see Fig. 6) that shows a perfect match with the node partition shown in Fig. 4.

C. Word associations

As a final example, let us use the University of South Florida Free Association Norms data set [28] to create a simple network[3] in the manner of [14]. We obtain a link partition by optimizing the modularity for the weighted line graph $\mathbf{D}$ of Eq. (11) but where the null model term $(k_\alpha k_\beta)/W^2$ has been scaled by a factor of 10.0 in order to control the resolution [9] and in this case obtain 321 communities in the whole network. The corresponding quality function can be seen as a linear approximation of the stability $R(\mathbf{D},n)$ [18]. It is easier to optimize for large networks.

In Fig. 7 we show part of the network near the word "bright" which is part of 11 communities[4]. The topology of our communities is much less constrained than those of k-clique percolation [14], which means we can pick out a wider range of structures. There are some tight cliquelike subsets, e.g., the names of the planets. At the other extreme the method finds more treelike structures such as the sequence "lit-on-switch-lever-handle," which is the backbone of another community linked to bright. On the other hand this flexibility in the structure can produce a confusing picture since many words are members of several communities though mostly having just one or two links per community. For instance for the word "bright," it is linked to eight of its 11 communities by just one link. However one can exploit this feature to start to define strength of membership in different communities. For instance for visualization, we have found it useful to view only those words that have a large number of links within one community, as in Fig. 7.

V. DISCUSSION

When describing a network, there seems to be a natural tendency to put the emphasis on its nodes whereas a graph is both a set of nodes and a set of links. It is therefore not surprising that node partitioning has been studied extensively in recent years while link partitioning has been overlooked so far. In this paper, we have shown that the quality of a link

[3]We take the sum of the two forward strengths of all pairs of normed word and add a link only if the total is greater than 0.025. We end up with 5018 words connected by 58 536 links and from this a line graph with 1 266 910 links is created.

[4]The 11 communities that contain "bright" are well characterized by the following subsets of words: ("brave," "bold," "daring"), ("bright," "light," "sunshine"), ("gone," "fade," "dim"), ("power," "electric," "lightning," "flash"), ("brain," "intelligence," "brilliant"), ("great," "wonderful," "gifted"), ("pen," "paper," "highlight"), ("handle," "lit," "on," "switch," "lever"), ("cloudy," "gray," "shiny," "sunny"), ("space," "sky," "moonlight," "stars"), ("assume," "illusion," "imagination," "vivid"). However "bright" has 16 of its 29 links in the community containing "sunshine" and "light" with just a single link to eight of its 11 communities.

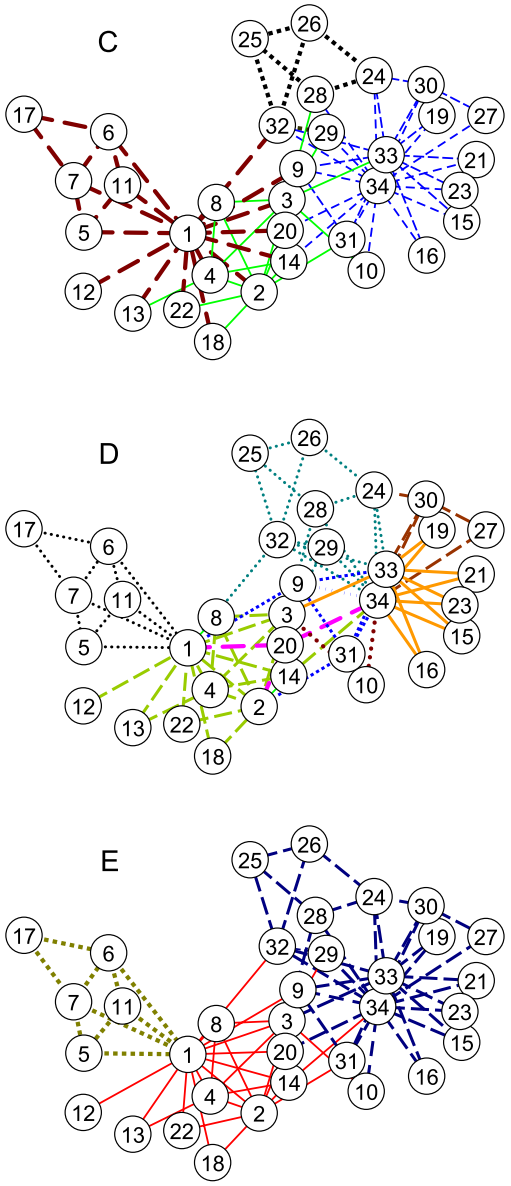

FIG. 5. (Color online) The optimal link partitions of (c) $Q(\mathbf{C})$, (d) $Q(\mathbf{D})$, and (e) $Q(E_1)$ for the Karate club. They contain four, seven, and three communities, respectively. The two smallest communities in the center of (d) consist of the links: (a) {(3,10), (10,34)}, (b) {(0,20), (1,20), (2,20)}.

partition can be evaluated by the modularity of its corresponding line graph. We have highlighted that optimizing the modularity of some of our weighted line graphs uncovers meaningful link partitions. Our approach has several advantages. A key criticism of the popular node partitioning methods is that a node must be in one single community whereas it is often more appropriate to attribute a node to several different communities. Link partitioning overcomes this limitation in a natural way. Moreover, the equivalence of a link partition of a graph G with the node partitioning of the corresponding line graph $L(G)$ means that one can use an existing node partitioning code with only the expense of producing a line graph transformation and an $O(\langle k^2\rangle/\langle k\rangle)$ increase in memory to accommodate the larger line graph.

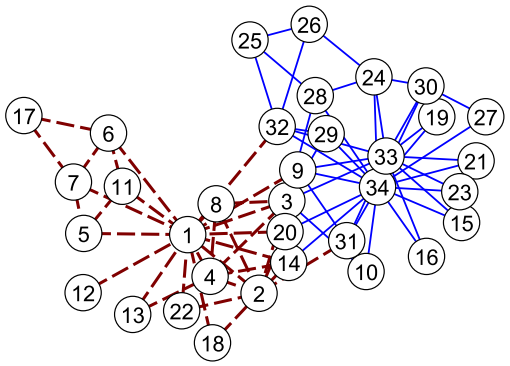

FIG. 6. (Color online) Optimal link partition into two communities of the stability $R(\mathbf{D},10)$ of the Karate club.

Even the memory cost can be reduced to be $O(1)$ since we have shown our link partitioning is equivalent to a process occurring on the links of the original graph G, so a line graph need not be produced explicitly.

Our method can be seen as a generalization of the popular k-clique percolation [14], which finds sets of connected k cliques. By way of comparison we find collections of two cliques, which are more densely connected than expected in an equivalent null model. Thus the link partitioning of our paper can be seen as an extension of two-clique percolation that allows for the uncovering of finer modules, i.e., two-clique percolation trivially uncovers connected components. An interesting generalization would be to apply our approach to the case of triangles, four cliques, etc. To do so, one has to replace the incidence matrix (relating nodes and links) by a more general bipartite graph, representing the membership of

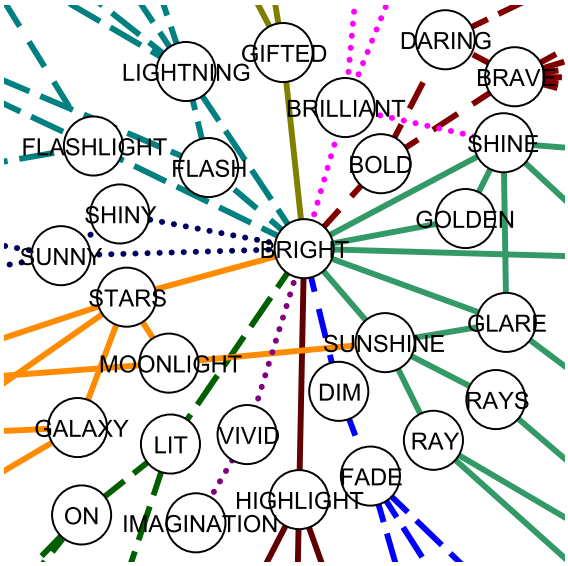

FIG. 7. (Color online) The simple graph created from the South Florida Free Association Norms data [28], in the manner of [14]. The link partition shown is produced by optimizing a modified version of the modularity $Q(\mathbf{D})$ where the null model factor was $10.0\times(k_\alpha k_\beta)/W^2$. This controls the number of communities found [9]. The subgraph shown contains the word "bright" along with nodes that have at least 90% of their links in one of the communities connected to "bright."

nodes in a clique of interest. Our random walk analysis in terms of this bipartite graph would then proceed in the same fashion and should allow to uncover finer modules than those obtained by k-clique percolation.

All our expressions also hold for the case of weighted networks. Even multiedges can be accommodated if we start from the incidence matrix, $\mathbf{B}$. However the beauty of our approach is that any type of graph analysis, be it community detection or something else, can be applied to a line graph rather than the original graph. In this way, one can view a network from a completely different angle yet use well established techniques to obtain fresh information about its structure.

ACKNOWLEDGMENTS

R.L. would like to thank M. Barahona and V. Eguiluz for interesting discussions and EPSRC-GB for support. After this work was finished, we received the paper of Ahn *et al.* [27] who also look at edge partitions but not in terms of weighted line graphs.

[1] M. Fiedler, Czech. Math. J. **25**, 619 (1975).
[2] W. Zachary, J. Anthropol. Res. **33**, 452 (1977).
[3] S. Fortunato and C. Castellano, in *Encyclopedia of Complexity and System Science*, edited by R. A. Meyers (Springer-Verlag, Berlin, 2009).
[4] M. A. Porter, J.-P. Onnela, and P. J. Mucha, e-print arXiv:0902.3788.
[5] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[6] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral, Phys. Rev. E **70**, 025101(R) (2004).
[7] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech. (2008) P10008.
[8] A. Noack and R. Rotta, Lect. Notes Comput. Sci. **5526**, 257 (2009).
[9] J. Reichardt and S. Bornholdt, Phys. Rev. E **74**, 016110 (2006).
[10] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[11] M. E. J. Newman, Phys. Rev. E **64**, 016131 (2001).
[12] R. S. Burt, Am. J. Sociol. **110**, 349 (2004).
[13] R. Lambiotte and P. Panzarasa, J. Informetrics **3**, 180 (2009).
[14] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).
[15] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri, J. Stat. Mech. (2009) P03024.
[16] A. Lancichinetti, S. Fortunato, and J. Kertész, New J. Phys. **11**, 033015 (2009).
[17] J.-C. Delvenne, S. Yaliraki, and M. Barahona, e-print arXiv:0812.1811.
[18] R. Lambiotte, J.-C. Delvenne, and M. Barahona, e-print arXiv:0812.1770.
[19] F. R. K. Chung, Spectral Graph Theory, CBMS Regional Conference Series in Mathematics.
[20] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[21] T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang, Phys. Rev. E **76**, 046115 (2007).
[22] R. Lambiotte and M. Ausloos, Phys. Rev. E **72**, 066107 (2005).
[23] V. K. Balakrishnan, *Schaum's Outline of Graph Theory* (Mcgraw-Hill Publishing Company, New York, 1997).
[24] H. Whitney, Am. J. Math. **54**, 150 (1932).
[25] L. R. Ford and D. R. Fulkerson, Can. J. Math. **8**, 399 (1956).
[26] G. Agarwal and D. Kempe, Eur. Phys. J. B **66**, 409 (2008).
[27] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, e-print arXiv:0903.3178.
[28] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber, The University of South Florida, word association, rhyme, and word fragment norms, http://www.usf.edu/FreeAssociation/
[29] A. Arenas, A. Fernandez, and S. Gomez, New J. Phys. **10**, 053039 (2008).