Network Science **3** (4): 445–479, 2015. © Cambridge University Press 2015
doi:10.1017/nws.2015.23

Relative centrality and local community detection

CHENG-SHANG CHANG, CHIH-JUNG CHANG, WEN-TING HSIEH, DUAN-SHIN LEE and LI-HENG LIOU

Institute of Communications Engineering, National Tsing Hua University, Hsinchu 300, Taiwan, Republic of China
(e-mail: cschang@ee.nthu.edu.tw, u9661234@oz.nthu.edu.tw, athena41393@livemail.tw, dacapo1142@gmail.com, lds@cs.nthu.edu.tw*)*

WANJIUN LIAO

Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, Republic of China
(e-mail: wjliao@ntu.edu.tw*)*

Abstract

In this paper, we develop a formal framework for what a good community should look like and how strong a community is (community strength). One of the key innovations is to incorporate the concept of relative centrality into structural analysis of networks. In our framework, relative centrality is a measure that measures how important a set of nodes in a network is with respect to another set of nodes, and it is a generalization of centrality. Building on top of relative centrality, the community strength for a set of nodes is measured by the difference between its relative centrality with respect to itself and its centrality. A community is then a set of nodes with a nonnegative community strength. We show that our community strength is related to conductance that is commonly used for measuring the strength of a small community. We define the modularity for a partition of a network as the average community strength for a randomly selected node. Such a definition generalizes the original Newman's modularity and recovers the stability in as special cases. For the local community detection problem, we also develop efficient agglomerative algorithms that guarantee the community strength of the detected local community.

Keywords: *network science, local community detection, centrality, modularity, clustering algorithms*

1 Introduction

Network science, as an emerging field of research, has received a lot of attention lately from many researchers in various fields, including physicists, mathematicians, computer scientists, biologists, and sociologists. As these researchers are from different fields and published their papers in different areas of journals, the terminology and the focus of their works could be quite different. In particular, physicists put more emphasis on the discovery of new physical insights such as phase transition and percolation, while mathematicians stress the importance of

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

the mathematical proofs for these newly discovered phenomenons. For computer scientists, the efficiency, in terms of computational complexity, of the algorithms that lead to these new findings is the main interest of their research. On the other hand, biologists and sociologists apply the algorithms to the networks from their data and seek meaningful explanations for their findings. In order for researchers from various research disciplines to carry out their scientific research, it is thus essential to have a *foundation* for network science (Brandes et al., 2013).

Laying a foundation for network science on top of a huge amount of papers published in the area of network analysis is a challenging task. The main difficulty is that researchers from different fields have different viewpoints. In particular, for the community detection problem (or the graph clustering problem), people have different opinions on what a *good community* should look like. There are several widely accepted notions for this: (i) a good community should have more edges within the community than the edges going outside the community (see e.g., Radicchi et al. (2004); Hu et al. (2008)) and thus *conductance* might be a good measure (see e.g., Andersen et al. (2006); Andersen & Lang (2006); Leskovec et al. (2008)), (ii) a good community should be well connected and thus it should be dense (Fortunato, 2010) and have a high clustering coefficient (Watts & Strogatz, 1998) or lots of k-cliques (Palla et al., 2005), (iii) a graph with a good community structure should behave quite differently from random graphs and thus modularity and the null model (Newman, 2004; Mucha et al., 2010) can be used for measuring how well a graph is partitioned, (iv) a good community should be cohesive and cannot be easily divided into disconnected components (Karrer et al., 2008; Leskovec et al., 2010; Yang & Leskovec, 2012), (v) a good community should have a high probability to trap a random walker inside the community for a certain period of time and thus one can use either data compression techniques for describing the path of the random walker inside the community (Rosvall & Bergstrom, 2007, 2008) or stability (Lambiotte, 2010; Delvenne et al., 2010) for measuring how well a graph is partitioned, and (vi) rumors are spread fast within a good community (Boyd et al., 2005). Based on these viewpoints, many algorithms (see the review papers in Fortunato (2010) and Porter et al. (2009)) have been developed in the literature and they might be classified as follows: (i) divisive algorithms (betweenness, spectral partitioning, sweep) (Newman & Girvan, 2004; Radicchi et al., 2004; Wu & Huberman, 2004; Duch & Arenas, 2005; Raghavan et al., 2007; Yang & Leskovec, 2012), (ii) agglomerative algorithms (Newman, 2004; Clauset et al., 2004; Lancichinetti et al., 2009), (iii) statistic and machine learning methods (spectral learning (Kamvar et al., 2003), kernel-based clustering algorithms (Dhillon et al., 2004; Kulis et al., 2009), exponential families (Long et al., 2007; Karrer & Newman, 2011)), (iv) data compression algorithms (Rosvall & Bergstrom, 2007, 2008) and (v) clique percolation methods (Palla et al., 2005). Various comparison studies of these algorithms can be found in Danon et al. (2005); Lancichinetti & Fortunato (2009); Leskovec et al. (2010).

Unlike the recent efforts in Yang & Leskovec (2012) for defining communities based on various metrics of *ground-truth*, in this paper we make an attempt to develop a *formal framework* for what a good community should look like and how strong a community is (community strength). Through this framework we can not only provide more physical insights and unified interpretations for various structural analyses of networks such as centralities and communities, but also develop efficient

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

algorithms for local community detection with guaranteed community strength. Our view of a good community, more like what it literally means in English, is

> *a group of people (nodes) who consider themselves much more important to themselves than to random people on the street.*

In social network analysis, centralities (Freeman, 1977, 1979; Newman, 2009) have been widely used for ranking the importance of nodes. For instance, movie stars who have a lot of fans can be easily identified as important nodes in social networks by using the degree centrality. However, we generally do not consider movie stars important persons *to us*. On the contrary, family members or friends are much more important *to us*. In view of this, we extend the concept of *centrality* and incorporate the concept of *relative centrality* into structural analysis of networks. In our framework, relative centrality is a (probability) measure that measures how important a set of nodes in a network is with respect to another set of nodes, and it is a generalization of *centrality* that only measures (or ranks) how important a set of nodes in a network is with respect to *the whole set of nodes in the network*. A set (of nodes) that has a much larger relative centrality with respect to itself than its centrality can thus be viewed as a *community*.

As mentioned before, people have different views. As such, relative centrality can only be formally defined on top of a specific viewpoint. To obtain a viewpoint of a network, one typical method is to "sample" the network, e.g., edge sampling, random walks, diffusion (Mucha et al., 2010), or random gossiping (Boyd et al., 2005). Mathematically, each sampling method renders a (probability) measure for a network that enables us to carry out further analysis. As in the probabilistic framework in Chang et al. (2011), we model a network as a graph G and "sample" the graph to generate a bivariate distribution $p(\cdot, \cdot)$ for a pair of two nodes. The bivariate distribution can be viewed as a normalized *similarity* measure (Liben-Nowell & Kleinberg, 2003) between a pair of two nodes. A graph G associated with a bivariate distribution $p(\cdot, \cdot)$ is then called a *sampled* graph. In this paper, we only consider the case that the bivariate distribution is *symmetric*. Under this assumption, the two marginal distributions of the bivariate distribution are the same and they represent the probability that a particular node is selected in the sampled graph. As such, the marginal distribution can be used for defining the *centrality* of a set as the probability that a selected node is in this set. The larger the centrality of a node is, the larger probability the node is selected. We show that our centrality measure recovers various centrality measures in the literature as special cases, including the degree centrality and the Katz centrality (Katz, 1953), as they simply correspond to various methods of sampling a graph.

As an extension of centrality, the concept of relative centrality of a set of nodes S_1 with respect to another set of nodes S_2 in our framework is formally defined as the *conditional* probability that one node of the selected pair of two nodes is in the set S_1 given that the other node is in the set S_2. When the set S_2 is taken to be the whole set of nodes in the graph, then the relative centrality of S_1 with respect to S_2 is simply reduced to the centrality of S_1, which is the probability that one node of the selected pair of two nodes is in the set S_1. We note that Bell (Bell, 2012) recently developed an independent framework for relative centrality that is quite different from ours. In our framework, the domain for a relative centrality

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

measure is the set of all the nodes in a network, while the domain of a local centrality measure in Bell (2012) is only restricted to a certain subset of nodes in a network.

Since our view for a community is a group of people (nodes) who consider themselves much more important to themselves than to random people on the street, the *community strength* for a set of nodes S is defined as the difference between its relative centrality with respect to itself and its centrality. Moreover, a set of nodes with a *non-negative* community strength is called a *community*. With such a definition of community, we are able to prove several mathematically equivalent statements of a *community* that can be intuitively explained by their social meanings. There are also several interesting interpretations for the community strength. In particular, we show that our community strength for a certain sampled graph is related to *conductance* that is commonly used for measuring the strength of a small community (Andersen et al., 2006; Andersen & Lang, 2006; Leskovec et al., 2008). Also, for a certain sampled graph, it is related to the probability that a random walker is trapped inside the community for a certain period of time (Rosvall & Bergstrom, 2007, 2008).

The original modularity in Newman & Girvan (2004) is a measure to quantify the strength of community structure in a partition of a graph and such a measure has been widely accepted for analyzing community structure in the literature. One of the well-known problems of Newman's modularity is its resolution limit in detecting communities (Fortunato & Barthelemy, 2007). As such, there are other measures, such as stability in Lambiotte (2010); Delvenne et al. (2010), that were proposed for quantifying the strength of community structure in a partition of a graph. However, when the communities in a network span a wide range of sizes, it is not possible to find an optimal value of the resolution parameter to identify simultaneously all the communities in a network (Lancichinetti & Fortunato, 2011; Granell et al., 2012). In our framework, we define the *modularity* for a partition of a sampled graph as the average community strength of the community to which a randomly selected nodes belong. As such, a high modularity for a partition of a graph implies that there are communities with strong community strengths. We then show that the original modularity in Newman & Girvan (2004) and stability in Lambiotte (2010); Delvenne et al. (2010) are special cases of our modularity for certain sampled graphs.

Based on our framework, we also develop efficient algorithms for local community detection. The problem of local community detection for a seed node w_0 is generally formulated as a problem to find a *community* S that contains the seed node w_0 (Clauset, 2005). There have been many methods proposed in the literature for the local community detection problem. One approach for local community detection is to map a graph into a Markov chain (either by a random walk or by diffusion). Since a Markov chain converges to its steady state exponentially fast (Diaconis & Stroock, 1991), one can deduce that there is a cut of the Markov chain if the Markov chain (mapped from the graph) does not converge as fast as expected (the mixed or cut lemma in Spielman & Teng (2004)). Such a cut can then be found by performing a sweep over an ordered transient state probabilities (Andersen et al., 2006; Andersen & Lang, 2006; Yang & Leskovec, 2012). The main difficulty of such an approach is the need to compute the transient state probabilities as it requires the knowledge of

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

the transition probability matrix of the Markov chain and thus the adjacency matrix of the graph. As such, local community detection cannot be done *locally*. Another approach is the agglomerative approach (Clauset, 2005; Lancichinetti et al., 2009; Xu et al., 2012; Huang et al., 2013) that repeatedly adds a "relatively important" node to the community until the maximum size of the community is reached. An agglomerative approach for local community detection is more computationally efficient as such an approach only needs to explore the network *locally*. Since the concept of relative centrality can be formally defined for a sampled graph in our framework, the agglomerative approach for local community detection can be generalized and mathematically scrutinized on top of our foundation. For a sampled graph, we then propose local community detection algorithms that can guarantee the community strength of the detected local community. There are two key features of our local community detection algorithms: (i) local search: only nodes in the *neighboring set* need to be explored, and (ii) recursive update: the relative centralities can be efficiently updated by using recursive formulae. The computational complexity of our algorithms is $O(k_{1,\max} \cdot k_{2,\max} \cdot (s_{\max})^2)$, where $s_{\max}$ is the maximum size of the community, $k_{1,\max}$ is the computational complexity for computing relative centrality for a pair of two nodes, and $k_{2,\max}$ is the maximum size of a neighboring set. This is of the same order as that in Clauset (2005) (when $k_{1,\max}$ is a constant). Unlike the agglomerative approach in Clauset (2005); Lancichinetti et al. (2009); Xu et al. (2012); Huang et al. (2013) that only works for a specific objective function, our local community detection algorithm has the freedom to choose the "viewpoint" to sample a graph. In particular, we consider sampling a graph with a random walk that has a path length not greater than 2. Such a sampling method allows us to look beyond the first neighbors so that we can take the clustering coefficient into account. We test such a sampling method on the dataset of the American football games (Girvan & Newman, 2002). Our experimental results show that those conferences with strong community strengths can be detected with 100% precision and 100% recall. On the other hand, for those conferences with very weak community strengths, our algorithm does not achieve high precision and recall.

Though our framework is built upon that in Chang et al. (2011), the new contributions of this paper are: (i) the new notion of relative centrality and its associated theory, and (ii) a local community detection algorithm based on the notion of relative centrality. Through this framework, we show that *various* existing notions and methods in network analysis, such as centralities, conductance, modularity, and stability, can have unified interpretations by choosing appropriate bivariate distributions. However, this does not mean every existing notion for network analysis can be unified in this framework. In particular, the betweenness centralities, defined as the number of geodesic paths traversing through nodes, cannot be directly applied in our framework.

The rest of the paper is organized as follows. In Section 2, we introduce the concept of relative centrality. Based on relative centrality, we then define the community strength and modularity in Section 3 and address their connections to the previous works in the literature. We propose our local community detection algorithms in Section 4 and test these algorithms in Section 5. The paper is concluded in Section 6, where we address possible extensions of our work.

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

2 Relative centrality

In this section, we introduce the concept of relative centrality. Such a concept will be used as the basic building block for the whole paper.

2.1 Sampling a network

As mentioned in Section 1, relative centrality is defined on top of a specific viewpoint and a viewpoint of a network can be obtained by sampling the network. For this, we first briefly review the probabilistic framework in Chang et al. (2011) that describes how a network is sampled.

In the literature, an undirected network is commonly modeled by an undirected graph $G(V_g, E_g)$, where V_g denotes the set of vertices (nodes) in the graph and E_g denotes the set of edges (links) in the graph. Let $n = |V_g|$ be the number of vertices in the graph and index the n vertices from $1, 2, \ldots, n$. Then, the graph $G(V_g, E_g)$ can also be characterized by an $n \times n$ adjacency matrix A, where

$$A_{vw} = \begin{cases} 1, & \text{if vertices } v \text{ and } w \text{ are connected,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Let $m = |E_g|$ be the number of edges in the graph and k_v be the degree of vertex v. From the adjacency matrix, we then have

$$m = \frac{1}{2} \sum_{v=1}^{n} \sum_{w=1}^{n} A_{vw} \tag{2}$$

and

$$k_v = \sum_{w=1}^{n} A_{vw}. \tag{3}$$

The main idea of the probabilistic framework in Chang et al. (2011) is to *sample* a network by randomly selecting two nodes V and W according to a specific bivariate distribution $p(\cdot, \cdot)$. Specifically, for a network with the set of nodes $\{1, 2, \ldots, n\}$, we have

$$\mathsf{P}(V = v, W = w) = p(v, w). \tag{4}$$

Let $p_V(v)$ (resp. $p_W(w)$) be the marginal distribution of the random variable V (resp. W), i.e.,

$$p_V(v) = \sum_{w=1}^{n} p(v, w) \tag{5}$$

and

$$p_W(w) = \sum_{v=1}^{n} p(v, w). \tag{6}$$

If $p(v, w)$ is *symmetric*, then $p_V(v) = p_W(v)$ for all v, and $p_V(v)$ is the probability that a randomly selected node is v.

Definition 2.1 (***Sampled graph***)
A graph $G(V_g, E_g)$ that is sampled by randomly selecting two nodes V and W according to a specific bivariate distribution $p(\cdot, \cdot)$ in equation (4) is called a *sampled graph* and it is denoted by the two tuple $(G(V_g, E_g), p(\cdot, \cdot))$.

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

For a given graph $G(V_g, E_g)$, there are many methods of to generate sampled graphs by specifying the needed bivariate distributions. One particular method is to generate V and W as the two end nodes of a *uniformly* selected edge from the graph $G(V_g, E_g)$. For this edge sampling method, we have

$$\mathsf{P}(V = v, W = w) = \frac{1}{2m} A_{vw}, \tag{7}$$

where m is the number of edges in the graph $G(V_g, E_g)$.

A more general method is to generate the needed bivariate distribution by randomly selecting the two ends of a *path*. This can be done by randomly selecting a path via mapping the adjacency matrix. Specifically, one can first choose a (matrix) function f that maps an adjacency matrix A to another non-negative matrix $f(A)$. Then, one can define a bivariate distribution from $f(A)$ by

$$\mathsf{P}(V = v, W = w) = \frac{1}{||f(A)||_1} f(A)_{vw} \tag{8}$$

where $||f(A)||_1 = \sum_v \sum_w |f(A)_{vw}|$ is usual "entrywise" matrix norm of the matrix $f(A)$. As described in Liben-Nowell & Kleinberg (2003), this can also be viewed as a way to compute the "similarity" score between a pair of two nodes v and w. In fact, if there is a bounded similarity measure $sim(v, w)$ that gives a high score for a pair of two "similar" nodes v and w, then one can map that similarity measure to a bivariate distribution $p(v, w)$ as follows:

$$p(v, w) = \frac{sim(v, w) - \text{MINsim}}{\sum_{x=1}^{n} \sum_{y=1}^{n} \Big(sim(x, y) - \text{MINsim} \Big)} \tag{9}$$

where

$$\text{MINsim} = \min_{1 \leqslant x, y \leqslant n} sim(x, y) \tag{10}$$

is the minimum value of all the similarity scores. As such, one can view a sampled graph $(G(V_g, E_g), p(\cdot, \cdot))$ as a graph $G(V_g, E_g)$ associated with a *normalized* similarity measure $p(\cdot, \cdot)$.

2.2 *Definition and properties of relative centrality*

In social network analysis, centralities (Freeman, 1977, 1979; Newman, 2009) have been widely used for ranking the most important or central nodes in a network. Intuitively, an important node in a graph has a higher probability of being "spotted" than an arbitrary node. Since V and W are two randomly selected nodes in a sample graph $(G(V_g, E_g), p(\cdot, \cdot))$ via the bivariate distribution $p(\cdot, \cdot)$, it seems plausible to define the centrality of a node as the *probability* that a node is selected. This leads us to define the centrality of a set of nodes in a sampled graph in Definition 2.2 below.

Definition 2.2 (***Centrality***)
For a sampled graph $(G(V_g, E_g), p(\cdot, \cdot))$ with a *symmetric* bivariate distribution $p(\cdot, \cdot)$, the *centrality* of a set of nodes S, denoted by $C(S)$, is defined as the probability that a node in S is selected, i.e.,

$$C(S) = \mathsf{P}(V \in S) = \mathsf{P}(W \in S). \tag{11}$$

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

Note that

$$\mathsf{P}(W \in S) = \mathsf{P}(W \in S | V \in V_g). \tag{12}$$

Another way to interpret the centrality of a set of nodes S is the *conditional probability* that the randomly selected node W is inside S given that the random selected node V is inside the whole set of nodes in the graph. As $p(\cdot,\cdot)$ can be viewed as a normalized similarity measure, such an observation leads us to define the *relative centrality* of a set of nodes S_1 with respect to another set of nodes S_2 as the conditional probability that the randomly selected node W is inside S_1 given that the random selected node V is inside S_2. Intuitively, a node w is relatively important to a set of node S_2 if the node w has a high probability of being "spotted" from the set of nodes in S_2. Such a definition is formalized in Definition 2.3 below.

Definition 2.3 (***Relative centrality***)
For a sampled graph $(G(V_g, E_g), p(\cdot,\cdot))$ with a *symmetric* bivariate distribution $p(\cdot,\cdot)$, the *relative centrality* of a set of nodes S_1 with respect to another set of nodes S_2, denoted by $C(S_1|S_2)$, is defined as the conditional probability that the randomly selected node W is inside S_1 given that the random selected node V is inside S_2, i.e.,

$$C(S_1|S_2) = \mathsf{P}(W \in S_1 | V \in S_2). \tag{13}$$

In particular, when $S_2 = V_g$ is the set of all the nodes in the graph, the relative centrality of a set of nodes S_1 with respect to V_g is reduced to the *centrality* of the set of nodes S_1, i.e.,

$$C(S_1|V_g) = \mathsf{P}(W \in S_1 | V \in V_g) = \mathsf{P}(W \in S_1) = C(S_1). \tag{14}$$

Note that

$$\begin{aligned} C(S_1|S_2) = \mathsf{P}(W \in S_1 | V \in S_2) &= \frac{\mathsf{P}(V \in S_2, W \in S_1)}{\mathsf{P}(V \in S_2)} \\ &= \frac{\sum_{v \in S_2} \sum_{w \in S_1} p(v,w)}{\sum_{v \in S_2} p_V(v)} \end{aligned} \tag{15}$$

where $P_V(v)$ is the marginal distribution of V in equation (5). Also,

$$C(S_1) = \mathsf{P}(W \in S_1) = \sum_{w \in S_1} \mathsf{P}(W = w) = \sum_{w \in S_1} p_W(w) \tag{16}$$

where $P_W(w)$ is the marginal distribution of W in equation (6). Since we assume that $p(\cdot,\cdot)$ is symmetric, we also have

$$C(S_1) = \mathsf{P}(V \in S_1) = \sum_{w \in S_1} p_V(w). \tag{17}$$

In the following, we show several properties of relative centrality.

Proposition 2.1
For a sampled graph $(G(V_g, E_g), p(\cdot,\cdot))$ with a *symmetric* bivariate distribution $p(\cdot,\cdot)$, the following properties for the relative centrality defined in Definition 2.3 hold.

i. $0 \leqslant C(S_1|S_2) \leqslant 1$ and $0 \leqslant C(S_1) \leqslant 1$. Moreover, $C(V_g|S_2) = 1$ and $C(V_g) = 1$.
ii. (Additivity) If S_1 and S_1' are two disjoint sets, i.e., $S_1 \cap S_1'$ is an empty set, then $C(S_1 \cup S_1'|S_2) = C(S_1|S_2) + C(S_1'|S_2)$ and $C(S_1 \cup S_1') = C(S_1) + C(S_1')$.

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

iii. (Monotonicity) If S_1 is a subset of S_1', i.e., $S_1 \subset S_1'$, then $C(S_1|S_2) \leqslant C(S_1'|S_2)$ and $C(S_1) \leqslant C(S_1')$.

iv. (Reciprocity)

$$C(S_1)C(S_2|S_1) = C(S_2)C(S_1|S_2).$$

As a result, $C(S_1) \geqslant C(S_2)$ if and only if $C(S_2|S_1) \leqslant C(S_1|S_2)$.

Proof

Since the relative centrality is a conditional probability and the centrality is a probability, the properties in (i), (ii), and (iii) follow trivially from the property of probability measures.

(iv) From the symmetric property of $p(\cdot,\cdot)$, the definitions of relative centrality and centrality in equations (13) and (11), it follows that

$$\begin{aligned} C(S_1)C(S_2|S_1) &= \mathsf{P}(W \in S_1)\mathsf{P}(W \in S_2|V \in S_1) \\ &= \mathsf{P}(V \in S_1)\mathsf{P}(W \in S_2|V \in S_1) = \mathsf{P}(V \in S_1, W \in S_2) \\ &= \mathsf{P}(V \in S_2, W \in S_1). \end{aligned} \tag{18}$$

Similarly, we also have

$$C(S_2)C(S_1|S_2) = \mathsf{P}(V \in S_1, W \in S_2) = \mathsf{P}(V \in S_2, W \in S_1). \tag{19}$$

Thus,

$$C(S_1)C(S_2|S_1) = C(S_2)C(S_1|S_2).$$

□

2.3 Illustrating examples

In this section, we provide several illustrating examples for relative centrality, including the degree centrality, the Katz centrality, and the continuous-time random walk.

Example 2.1 (***Degree centrality***)

Consider the bivariate distribution in equation (7), i.e.,

$$p(v,w) = \mathsf{P}(V = v, W = w) = \frac{1}{2m}A_{vw}. \tag{20}$$

Thus, we have from equations (15) and (20) that

$$C(S_1|S_2) = \frac{\sum_{v\in S_2}\sum_{w\in S_1} A_{v,w}}{\sum_{v\in S_2} k_v}. \tag{21}$$

In particular, when S_1 contains a single node w and $S_2 = V_g$ is the set of all the nodes in the graph, the centrality of w, i.e., $C(\{w\})$, is simply $\mathsf{P}(W = w)$. From equation (20), we have

$$C(\{w\}) = \mathsf{P}(W = w) = \frac{k_w}{2m}. \tag{22}$$

Thus, $C(\{w\})$ is the usual (normalized) degree centrality that counts the number of edges connected to the node w.

As another illustrating example, we show that Katz's centrality can also be derived as a special case of the sampled graph.

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

Example 2.2 (***Katz centrality***)
For an undirected graph $G(V_g, E_g)$, Katz centrality (Katz, 1953) is based on the assumption that the importance of a vertex is the sum of a fixed constant and the "discounted" importance of vertices it connects to. Specifically, let x_v be the Katz centrality of vertex v. Then

$$x_v = \sum_{w=1}^{n} \lambda A_{vw} x_w + 1 \tag{23}$$

where λ is the discount factor. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ be the vector of Katz centralities and $\mathbf{1} = (1, 1, \ldots, 1)^T$ be the n-vector of all 1's. Then, we can write equation (23) in the matrix form $\mathbf{x} = \lambda A\mathbf{x} + \mathbf{1}$ and solve for $\mathbf{x}$ to yield

$$\mathbf{x} = (\mathbf{I} - \lambda A)^{-1} \cdot \mathbf{1} \tag{24}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix. To show that the Katz centrality is the centrality of a particular sampled graph, we choose

$$f(A) = \sum_{i=0}^{\infty} (\lambda A)^i = (\mathbf{I} - \lambda A)^{-1} \tag{25}$$

in equation (8) and thus we have

$$p(v, w) = \frac{1}{||(\mathbf{I} - \lambda A)^{-1}||_1} (\mathbf{I} - \lambda A)^{-1}. \tag{26}$$

The matrix $(\mathbf{I} - \lambda A)^{-1}$ in equation (25) is called the "Katz similarity" in Newman (2009) and the bivariate distribution in equation (26) can thus be viewed as the normalized Katz similarity. Since A is the adjacency matrix of an undirected graph, we have $A = A^T$ and thus $p(v, w)$ is symmetric. This then leads to

$$C(\{w\}) = p_W(w) = p_V(w) = \frac{\mathbf{e}_w^T}{||(\mathbf{I} - \lambda A)^{-1}||_1} (\mathbf{I} - \lambda A)^{-1} \cdot \mathbf{1} \tag{27}$$

where $\mathbf{e}_w$ is the column vector that has 1 in its wth element and 0 otherwise. Thus, $C(\{w\})$ is the (normalized) Katz centrality of node w. For this example, we also have from the symmetry of A, equations (15), and (26) that

$$C(S_1|S_2) = \frac{\mathbf{e}_{S_2}^T \cdot (\mathbf{I} - \lambda A)^{-1} \cdot \mathbf{e}_{S_1}}{\mathbf{e}_{S_2}^T \cdot (\mathbf{I} - \lambda A)^{-1} \cdot \mathbf{1}} \tag{28}$$

where $\mathbf{e}_S$ is the column vector that has 1 in the elements of the set S and 0 otherwise.

Example 2.3 (***Continuous-time random walk on an undirected graph***)
A continuous-time random walk on an undirected graph $G = (V_g, E_g)$ is a continuous-time Markov process with the transition rate from vertex v to vertex w being A_{vw}/k_v. Let $\gamma_v(t)$ be the probability that the continuous-time random walk is at vertex v at time t when the walk is started from vertex w at time 0. Since that the transition rate from vertex v to vertex w is A_{vw}/k_v, it then follows that

$$\frac{d\gamma_v(t)}{dt} = \sum_{w=1}^{n} \frac{A_{wv}}{k_w} \gamma_w(t) - \gamma_v(t) = \sum_{w=1}^{n} \frac{A_{vw}}{k_w} \gamma_w(t) - \gamma_v(t) \tag{29}$$

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

where we use $A = A^T$ in the last equality. Let π_v be the steady probability that the continuous-time random walk is at vertex v. In view of equation (29),

$$\pi_v = \lim_{t\to\infty} \gamma_v(t) = \frac{k_v}{2m}. \tag{30}$$

Let $\gamma(t) = (\gamma_1(t), \ldots, \gamma_n(t))^T$ and $L = (L_{v,w})$ be the normalized graph Laplacian, i.e.,

$$L_{v,w} = \frac{A_{vw}}{k_w} - \delta_{v,w}. \tag{31}$$

We can rewrite equation (29) in the matrix form as follows:

$$\frac{d\gamma(t)}{dt} = L\gamma(t). \tag{32}$$

This then leads to

$$\gamma(t) = e^{tL}\gamma(0). \tag{33}$$

Since we start from vertex w at time 0, i.e., $\gamma_w(0) = 1$, we then have

$$\gamma_v(t) = (e^{tL})_{vw}. \tag{34}$$

Now we choose the two nodes V and W as the two ends of a path of length t in a stationary continuous-time random walk. Specifically,

$$\begin{aligned} p(v,w) &= \mathsf{P}(V = v, W = w) = \mathsf{P}(W = w)\mathsf{P}(V = v|W = w) \\ &= \pi_w\gamma_v(t) = (e^{tL})_{vw}\frac{k_w}{2m} \end{aligned} \tag{35}$$

where we use equations (34) and (30) in the last identity. Since a stationary continuous-time random walk is a reversible Markov process, we also have

$$p(v,w) = \mathsf{P}(V = v, W = w) = \mathsf{P}(V = w, W = v) = p(w,v).$$

For this example, we then have

$$C(S_1|S_2) = \frac{\sum_{v\in S_2}\sum_{w\in S_1}(e^{tL})_{vw}\frac{k_w}{2m}}{\sum_{v\in S_2}\frac{k_v}{2m}} \tag{36}$$

$$C(S_1) = \sum_{w\in S_1}\frac{k_w}{2m}. \tag{37}$$

We note that our definition for relative centrality can also be applied to other centrality measures that are based on similarity measures, e.g., the harmonic mean closeness centrality of a node w (Newman, 2009) is defined as $\frac{1}{n-1}\sum_{v\neq w} 1/d_{v,w}$, where $d_{v,w}$ is the length of a geodesic path from vertex v to vertex w. However, betweenness centralities, defined as the number of geodesic paths traversing through nodes, cannot be directly applied.

3 Community strength and modularity

In this section, we define *community strength* and *modularity* based on relative centrality. We will show that our definition generalizes the original Newman's definition (Newman, 2004) and unifies various other generalizations, including stability in Lambiotte (2010); Delvenne et al. (2010).

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

3.1 Community strength

Intuitively, a community in a social network includes a group of people who consider themselves much more important to themselves than to random people on the street. In view of this, one might consider a community as a group of nodes that have high relative centralities to each other than to random nodes. As such, we propose the following definition of *community strength* based on relative centrality.

Definition 3.1 (***Community strength***)
For a sample graph $(G(V_g, E_g), p(\cdot,\cdot))$ with a symmetric bivariate distribution $p(\cdot,\cdot)$, the *community strength* of a subset set of nodes $S \subset V_g$, denoted by $Str(S)$, is defined as the difference of the relative centrality of S with respect to itself and its centrality, i.e.,

$$Str(S) = C(S|S) - C(S). \tag{38}$$

In particular, if a subset set of nodes $S \subset V_g$ has a non-negative community strength, i.e., $Str(S) \geqslant 0$, then it is called a *community*.

To understand the definition of the community strength, let us consider the continuous-time random walk in Example 2.3. Let $X(t)$ be the node that the continuous-time random walk visits at time t. Then

$$\begin{aligned} &C(S|S) - C(S) \\ &\quad = \mathsf{P}(W \in S | V \in S) - \mathsf{P}(W \in S) \\ &\quad = \mathsf{P}(X(t) \in S | X(0) \in S) - \mathsf{P}(X(t) \in S). \end{aligned} \tag{39}$$

The first term in equation (39) is the conditional probability that the continuous-time random walk is "trapped" in the set S given that it is started from that set, while the second term is the steady state probability that the random walk is in the set S. In view of equation (39), the continuous-time random walk is more likely being "trapped" in a set with a strong community strength. Of course, the large the set is, the more likely the random walk will be "trapped" in that set. As such, the community strength has to take into account the steady state probability that the random walk is in the set. By so doing, for the whole set of nodes in the graph, i.e., V_g, the community strength is normalized to 0 as it should not contain any information for the strength of this trivial community.

To further understand the physical meaning of the concept of community strength, we show how it is related to *conductance* for a small community. In the literature, conductance has been widely used for testing whether a community (cut) is good (see e.g., Andersen et al. (2006); Andersen & Lang (2006); Leskovec et al. (2008)) when the community (cut) is relatively small comparing to the whole graph. The conductance of a set S, denoted by $\phi(S)$, is defined as

$$\frac{\sum_{v\in S}\sum_{w\notin S} A_{v,w}}{\min[\sum_{v\in S} k_v, \sum_{v\notin S} k_v]}. \tag{40}$$

When S is relatively small comparing to the whole graph, we usually have

$$\sum_{v\in S} k_v \leqslant \sum_{v\notin S} k_v$$

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

and the conductance of S is reduced to

$$\phi(S) = \frac{\sum_{v\in S}\sum_{w\notin S} A_{v,w}}{\sum_{v\in S} k_v}. \tag{41}$$

In view of equation (41), a small community S with a small conductance can be viewed as a good community. Now let us consider the bivariate distribution in equation (7), where V and W represent the two ends of a uniformly selected edge. For this case, we have from equation (22) that

$$C(S) = \frac{\sum_{v\in S} k_v}{2m}$$

where m is the total number of edges. For a small community, we expect $\sum_{v\in S} k_v \ll m$ and $C(S) \approx 0$. Then, we have from equation (21) in Example 2.1 that

$$\begin{aligned} Str(S) \approx C(S|S) &= \frac{\sum_{v\in S}\sum_{w\in S} A_{v,w}}{\sum_{v\in S} k_v} \\ &= 1 - \frac{\sum_{v\in S}\sum_{w\notin S} A_{v,w}}{\sum_{v\in S} k_v} = 1 - \phi(S). \end{aligned} \tag{42}$$

In view of equation (42), a small community S with a large community strength has a small conductance and thus can be considered as a good community.

We note from equation (39) that the definition of a community, i.e., a set with a non-negative community strength, is a generalization of the definition of a community in Chang et al. (2011). The concept of *community strength* based on relative centrality enables us to look at the definition of a community from various perspectives. For instance, we know from equation (38) that a community is a set (of nodes) with its relative centrality to itself not less than its centrality. In addition to this, there are other various equivalent statements for a community that can be explained by their social meanings. These are as shown in Theorem 3.1 below. The proof of Theorem 3.1 is given in Appendix A.

Theorem 3.1

Consider a sample graph $(G(V_g, E_g), p(\cdot,\cdot))$ with a symmetric bivariate distribution $p(\cdot,\cdot)$, and a set S with $0 < C(S) < 1$. Let $S^c = V_g \backslash S$ be the set of nodes that are not in S. The following statements are equivalent.

i. The set S is a community, i.e., $Str(S) = C(S|S) - C(S) \geqslant 0$.
ii. The relative centrality of S with respect to S is not less than the relative centrality of S with respect to S^c, i.e., $C(S|S) \geqslant C(S|S^c)$.
iii. The relative centrality of S^c with respect to S is not greater than the centrality of S^c, i.e., $C(S^c|S) \leqslant C(S^c)$.
iv. The relative centrality of S with respect to S^c is not greater than the centrality of S, i.e., $C(S|S^c) \leqslant C(S)$.

As mentioned before, the social meaning for the first statement in Theorem 3.1(i) is that a community is a group of people who consider themselves much more important to themselves than to random people on the street. The second statement in Theorem 3.1(ii) says that a community is a group of people who consider themselves much more important to themselves than to the other people not in the community. The third statement in Theorem 3.1(iii) says that the other people not

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

in a community are much less important to the people in the community than to random people on the street. Finally, the fourth statement in Theorem 3.1(iv) says that people in a community are much less important to the other people not in the community than to random people on the street.

3.2 Modularity

In Newman (2004), Newman proposed using *modularity* as a metric that measures the quality of a division of a network from the global perspective. Such an index has been widely accepted for analyzing community structure in the literature. By using community strength, we propose a (generalized) modularity index and our index will be shown to recover the original Newman's modularity and stability in Lambiotte (2010); Delvenne et al. (2010) as special cases. Since community strength is a metric that measures the quality of the structure of a community from the perspective of the nodes inside the community, it seems reasonable to define modularity as the average community strength of the community to which a randomly selected nodes belong.

Definition 3.2 (***Modularity***)
Consider a sampled graph $(G(V_g, E_g), p(\cdot,\cdot))$ with a symmetric bivariate distribution $p(\cdot,\cdot)$. Let S_c, $c = 1, 2, \ldots, C$, be a partition of $\{1, 2, \ldots, n\}$, i.e., $S_c \cap S_{c'}$ is an empty set for $c \neq c'$ and $\cup_{c=1}^{C} S_c = \{1, 2, \ldots, n\}$. The modularity index Q with respect to the partition S_c, $c = 1, 2, \ldots, C$, is defined as the weighted average of the community strength of each subset with the weight being the centrality of each subset, i.e.,

$$Q = \sum_{c=1}^{C} C(S_c) \cdot Str(S_c). \tag{43}$$

As the centrality of a set S_c is the probability that a randomly selected node is in the set S_c, the modularity index Q in equation (43) is the average community strength of the community to which a randomly selected node belongs. Such a definition is also the same as that in Chang et al. (2011) as

$$\sum_{c=1}^{C} C(S_c) \cdot Str(S_c) = \sum_{c=1}^{C} \Big(\mathsf{P}(V \in S_c, W \in S_c) - \mathsf{P}(V \in S_c)\mathsf{P}(W \in S_c) \Big).$$

It was shown in Chang et al. (2011) that the (generalized) modularity in Definition 3.2 is in fact a generalization of the original modularity index in Newman (2004) by choosing the bivariate distribution $p(v, w)$ in equation (7), i.e.,

$$\mathsf{P}(V = v, W = w) = \frac{1}{2m} A_{vw}. \tag{44}$$

For such a choice, the modularity index Q in equation (43) is

$$\sum_{c=1}^{C} \sum_{v \in S_c} \sum_{w \in S_c} \Big(\frac{1}{2m} A_{vw} - \frac{k_v}{2m} \frac{k_w}{2m} \Big). \tag{45}$$

One of the well-known problems of using Newman's modularity in equation (45) is its resolution limit in detecting communities (Fortunato & Barthelemy, 2007)

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

smaller than a certain scale. This motivated many researchers to propose multi-scale methods, including the *stability* in Lambiotte (2010); Delvenne et al. (2010). In the following example, we show that stability can also be derived as a special case of a sampled graph.

Example 3.1 (***Stability in Lambiotte (2010); Delvenne et al. (2010)***)
If we choose the bivariate distribution $p(v, w)$ as in equation (35), i.e.,

$$p(v,w) = p(w,v) = (e^{tL})_{vw}\pi_w = (e^{tL})_{vw}\frac{k_w}{2m} \tag{46}$$

then the modularity index Q in equation(43) is simply the stability previously defined in Lambiotte (2010); Delvenne et al. (2010), i.e.,

$$\sum_{c=1}^{C}\sum_{v\in S_c}\sum_{w\in S_c}\left((e^{tL})_{vw}\frac{k_w}{2m} - \frac{k_v}{2m}\frac{k_w}{2m}\right). \tag{47}$$

Since V and W in this example are the two ends of a randomly selected path via a continuous-time random walk with time t, the parameter t serves as a multi-scale resolution parameter for the size of the communities. Intuitively, a large (resp. small) resolution parameter t tends to identify large (resp. small) communities. Also, it is further illustrated in Lambiotte (2010); Delvenne et al. (2010) that stability is in fact a generalization of other multi-scale methods in Reichardt & Bornholdt (2006a); Arenas et al. (2008).

4 Local community detection

The problem of local community detection for a seed node w_0 is generally formulated as a problem to find a *community* S that contains the seed node w_0. There have been many methods proposed in the literature for the local community detection problem (see e.g., Spielman & Teng (2004); Palla et al. (2005); Clauset (2005); Andersen et al. (2006); Andersen & Lang (2006); Lancichinetti et al. (2009); Yang & Leskovec (2012); Huang et al. (2013)).

Our approach for local community detection is an agglomerative approach like those in Clauset (2005); Lancichinetti et al. (2009); Huang et al. (2013). We start from a seed node and repeatedly add a node to the local community until either the size of community is reached or no more nodes can be added. The criterion of choosing the added node is to make sure the community strength can be maintained and the added node has the largest relative centrality with respect to the current community. One crucial point of this approach, as pointed out in Clauset (2005), is to explore the graph *locally*. Typically, this is done by visiting neighboring nodes. In the following section, we introduce the notions of *positively correlated sets* and *neighboring sets* that allow us to explore a sample graph locally.

4.1 Positively correlated sets and neighboring sets

The concept of positively correlated sets are quite similar to two groups of people who are friendly to each other, i.e., they are nicer to each other than to random people on the street. Such a concept is formally defined below.

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

Definition 4.1
Consider a sampled graph $(G(V_g, E_g), p(\cdot,\cdot))$ with a symmetric bivariate distribution $p(\cdot,\cdot)$. Two sets of nodes S_1 and S_2 are said to be *positively correlated* if the relative centrality of S_1 with respect to S_2 is not smaller than the centrality of S_1, i.e.,

$$C(S_1|S_2) \geqslant C(S_1). \tag{48}$$

Note from the reciprocity in Proposition 2.1(iv) that equation (48) is equivalent to

$$C(S_2|S_1) \geqslant C(S_2). \tag{49}$$

In order to maintain the community strength of a *growing* community, it is intuitive to only invite people who are *friendly* to the current community to join. In the following lemma, we show that such an intuition can be formally proved by merging two positively correlated and disjoint sets. We show the new set has the community strength not less than the minimum of the community strengths of these two sets. This key lemma is an important step in our agglomerative approach as it enables us to add a positively correlated node to a community without weakening its community strength. The proof of Lemma 4.1 is given in Appendix B.

Lemma 4.1
Consider a sampled graph $(G(V_g, E_g), p(\cdot,\cdot))$ with a symmetric bivariate distribution $p(\cdot,\cdot)$. Suppose that S_1 and S_2 are two disjoint sets.

i. For an arbitrary set S_3, the relative centrality of S_3 with respect to $S_1 \cup S_2$ can be computed as follows:

$$C(S_3|S_1 \cup S_2) = \frac{C(S_1)\cdot C(S_3|S_1) + C(S_2)\cdot C(S_3|S_2)}{C(S_1) + C(S_2)}. \tag{50}$$

ii. The community strength of $S_1 \cup S_2$ can be computed as follows:

$$Str(S_1 \cup S_2) = \frac{C(S_1)\Big(Str(S_1) + 2(C(S_2|S_1) - C(S_2))\Big)}{C(S_1) + C(S_2)} + \frac{C(S_2)\cdot Str(S_2)}{C(S_1) + C(S_2)}. \tag{51}$$

If, furthermore, S_1 and S_2 are positively correlated, then

$$Str(S_1 \cup S_2) \geqslant \min[Str(S_1), Str(S_2)]. \tag{52}$$

In order to explore a sampled graph locally, we introduce the notion of *neighboring sets* below.

Definition 4.2 (***Neighboring sets***)
The neighboring set of a set of nodes S, denoted by $Nei(S)$, is the set of nodes that are not in S and has positive relative centrality with respect to S, i.e.,

$$Nei(S) = \{w : w \notin S,\ C(\{w\}|S) > 0\}. \tag{53}$$

In the following proposition, we show that only nodes in the neighboring set of a set S can be positively correlated to S. As such, if we would like find nodes that are positively correlated to a set S, we only need to explore the nodes in its neighboring set. Furthermore, we also show how the neighboring set of a set S is updated when a new node is added to S. The proof of Proposition 4.1 is given in Appendix C.

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

Proposition 4.1
Suppose that $C(S) > 0$ for any nonempty set S.

i. Any node w that is not in $Nei(S) \cup S$ cannot be positively correlated to S.
ii. Suppose that a node w^* is not in S. Then

$$Nei(S \cup \{w^*\}) = \Big(Nei(S) \backslash \{w^*\}\Big) \cup \Big(Nei(\{w^*\}) \backslash (S \cup Nei(S))\Big). \tag{54}$$

iii. Suppose that a node w^* is not in S. Then for a node w in $Nei(\{w^*\}) \backslash (S \cup Nei(S))$,

$$C(\{w\}|S \cup \{w^*\}) = \frac{C(\{w^*\}) \cdot C(\{w\}|\{w^*\})}{C(S) + C(\{w^*\})}. \tag{55}$$

4.2 Local community detection algorithm

In this section, we propose local community detection algorithms for sampled graphs. Given a specific node w_0 with community strength not less than γ, the algorithm generates a community that contains the node w_0 with community strength not less than γ. There are three ideas of the algorithm. Our first idea is to start from w_0 and recursively add a node that is both positively correlated to the local community and has the community strength not less than γ. According to Lemma 4.1(ii), such a merging procedure will yield a new community that contains w_0 and has community strength not less than γ. Such a merging procedure is repeated until either there are no more nodes that can be added or the community size is reached. To reduce the computational complexity, our second idea is to use Proposition 4.1 to limit the size of the nodes that need to be explored in every iteration. According to Proposition 4.1, only the neighboring set of the newly added nodes needs to be explored. Finally, we use Proposition 2.1 (ii) and Lemma 4.1(i) for recursive updates of the needed centralities and relative centralities.

Local community detection algorithm:
(P0) Input a sampled graph $(G(V_g, E_g), p(\cdot,\cdot))$ with three (external) functions: the first function $C_e(\{w\})$ with input w can be called to compute the centrality of node w, and the second function $C_e(\{w\}|\{v\})$ with inputs w and v can be called to compute the relative centrality of $\{w\}$ with respect to $\{v\}$, and the third function $Nei_e(\{w\})$ with input w can be called to compute the neighboring set of node w. Given a seed node w_0 with $Str(\{w_0\}) \geqslant \gamma$ and the maximum size $s_{\max}$ of the local community (with $s_{\max} \leqslant n$), the algorithm generates a community containing w_0 that has the community size not greater than $s_{\max}$ and the community strength not less than γ.
(P1) (Initialization) Initially, the local community S is an empty set and set $C(S) = 0$ and $Nei(S)$ to be an empty set. Set the chosen node w^* to be the seed node w_0, i.e, $w^* \leftarrow w_0$.
(P2) (Exploring new neighbors) Update the neighboring set by using equation (54) in Proposition 4.1(ii), i.e.,

$$Nei(S) \leftarrow (Nei(S) \backslash \{w^*\}) \cup \Big(Nei_e(\{w^*\}) \backslash (S \cup Nei(S))\Big).$$

For every node w in $Nei_e(\{w^*\}) \backslash (S \cup Nei(S))$, compute and store the relative centrality of $\{w\}$ with respect to S by using equation (55) in Proposition 4.1, i.e.,

$$C(\{w\}|S) \leftarrow \frac{C_e(\{w^*\}) \cdot C_e(\{w\}|\{w^*\})}{C(S) + C_e(\{w^*\})}. \tag{56}$$

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

Compute and store the centrality of $\{w\}$ by

$$C(\{w\}) \leftarrow C_e(\{w\}). \tag{57}$$

Compute and store the community strength of $\{w\}$ by

$$Str(\{w\}) \leftarrow C_e(\{w\}|\{w\}) - C_e(\{w\}). \tag{58}$$

Extend the local community S by adding w^* into the local community, i.e.,

$$S \leftarrow S \cup \{w^*\}.$$

Update the centrality of S by

$$C(S) \leftarrow C(S) + C_e(\{w^*\}).$$

(P3) (Finding candidates among neighbors) Find the set of *candidate* nodes S^+ in $Nei(S)$ that have community strength not less than γ and are positively correlated to S, i.e.,

$$S^+ = \{w : w \in Nei(S),\ Str(\{w\}) \geqslant \gamma,\ C(\{w\}|S) - C(\{w\}) > 0\}.$$

(P4) (Stopping criteria) If either the set of candidate nodes S^+ is empty or the maximum size $s_{\max}$ is reached, i.e., $|S| = s_{\max}$, return the set S as the detected local community. Otherwise, do (P5).

(P5) (Optimal selection) Choose the node w^* in S^+ that has the largest relative centrality with respect to S, i.e.,

$$w^* = \arg\max_{w \in S^+} C(\{w\}|S). \tag{59}$$

(P6) (Updating relative centralities) For every node w in $Nei(S)\backslash\{w^*\}$, update its relative centrality with respect to S by using the rule in equation (50), i.e.,

$$C(\{w\}|S) \leftarrow \frac{C(S) \cdot C(\{w\}|S) + C(\{w^*\}) \cdot C_e(\{w\}|\{w^*\})}{C(S) + C(\{w^*\})}. \tag{60}$$

(P7) Repeat (P2).

Theorem 4.1

The set S returned by the local community detection algorithm above has community strength not less than γ, i.e., $Str(S) \geqslant \gamma$. If particular, if $\gamma \geqslant 0$, then S is a community.

Proof

As the seed node w_0 has community strength not less than γ (as described in P(0)), the community strength of $S = \{w_0\}$ in (P2) is not less than γ. From (P3), we know that every node w in S^+ has community strength not less than γ and it is also positively correlated to S. As a result of equation (52) in Lemma 4.1, the set $S \cup \{w^*\}$ also has community strength not less than γ. Repeating the same argument shows that the set S returned by the local community detection algorithm has community strength not less than γ. □

Now we analyze the computational complexity of the above algorithm by assuming that all the three (external) functions take at most $k_{1,\max}$ steps and that there are at most $k_{2,\max}$ neighboring nodes of a node. Note from (P4) that there are at most $s_{\max}$ iterations in the above algorithm. In (P2), there are at most $k_{2,\max}$ new nodes that need to be explored. The computation for the relative centralities in equation (56) requires

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

$O(k_{1,\max})$ steps. Thus, the computational complexity for (P2) is $O(k_{1,\max} \cdot k_{2,\max})$. In each iteration, the computational complexity in (P3), (P5), and (P6) is proportional to the size of $Nei(S)$. Since there are at most $k_{2,\max}$ neighboring nodes of a node, the size of $Nei(S)$ is not greater than $k_{2,\max} \cdot s_{\max}$ and the computational complexity in (P3), (P5), and (P6) is $O(k_{1,\max} \cdot k_{2,\max} \cdot s_{\max})$ (as there is an external function call in (P6)). Thus, the overall computational complexity for the above algorithm is

$$O(k_{1,\max} \cdot k_{2,\max} \cdot (s_{\max})^2).$$

We note that our algorithm can be easily extended to the local community detection problem with a seed set S_0. To see this, suppose that the seed set S_0 contains k nodes $w_1, w_2, \ldots, w_k$. Then, we can simply set the chosen node w^* to be w_i in the ith iteration for $i = 1, 2 \ldots, k$. After that, we resume the process of selecting the optimal node in (P5).

5 Experiment results

In this section, we perform various experiments on real-world networks by using the local community detection algorithm in the previous section.

5.1 Selecting the bivariate distribution

Though there are many choices of the bivariate distribution $p(\cdot, \cdot)$ for computing relative centralities, not every one of them is suitable for the purpose of local community detection. For instance, the bivariate distributions in Examples 2.2 and 2.3 require the complete knowledge of the adjacency matrix A of the graph and they defeat the purpose of exploring the graph locally. On other hand, the degree centrality in Example 2.1 only requires the *local information* that contains the *first* neighbors of v and w for computing the needed relative centralities. However, exploring a graph by only looking at the first neighbors might lead to a very limited view that sometimes prohibits us to tell the difference among all the first neighbors. For example, as shown in Figure 1, all the three nodes w_1, w_2, and w_3 have the same relative centrality to the node v when the degree centrality in Example 2.1 is used. If we further take the *second* neighbors into consideration, we see that the three nodes w_1, w_2, and v forms a triangle. As such, it seems that w_1 and w_2 should have larger relative centralities with respect to v than that of w_3. This motivates us to use random walks with path length not greater than 2 for computing the relative centralities in our local community detection algorithm.

Example 5.1 (***A random walk with path length not greater than 2***)
A random walk with path length not greater than 2 can be generated by the following two steps: (i) with the probability $k_v/2m$, an initial node v is chosen, (ii) with probability β_i, $i = 0, 1, 2$, a walk from v to w with length i is chosen. As such, we have

$$p(v, w) = \frac{\beta_0 k_v \delta_{v,w}}{2m} + \frac{\beta_1}{2m} A_{v,w} + \frac{\beta_2}{2m} \sum_{v_2=1}^{n} \frac{A_{v,v_2} A_{v_2,w}}{k_{v_2}} \tag{61}$$

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

where $\beta_0 + \beta_1 + \beta_2 = 1$ and $\beta_i \geqslant 0$, $i = 1, 2, 3$. Thus,

$$C(\{w\}|\{v\}) = \mathsf{P}(W = w|V = v) = \frac{\left(\beta_0 k_v \delta_{v,w} + \beta_1 A_{v,w} + \beta_2 \sum_{v_2=1}^{n} \frac{A_{v,v_2} A_{v_2,w}}{k_{v_2}}\right)}{k_v} \tag{62}$$

and

$$C(\{w\}) = \frac{k_w}{2m}. \tag{63}$$

In view of equation (62), we know that $C(\{w\}|\{v\}) = 0$ if node w is neither a first neighbor nor a second neighbor of v. Thus, the neighboring set of a node v, $Nei(\{v\})$, is simply the union of its first neighbors and second neighbors.

Also, observe that

$$\begin{aligned} C(S|S) = \mathsf{P}(W \in S|V \in S) &= \frac{\mathsf{P}(W \in S, V \in S)}{\mathsf{P}(V \in S)} \\ &= \frac{\sum_{v\in S} \sum_{w\in S} \left(\beta_0 k_v \delta_{v,w} + \beta_1 A_{v,w} + \beta_2 \sum_{v_2=1}^{n} \frac{A_{v,v_2} A_{v_2,w}}{k_{v_2}}\right)}{\sum_{v\in S} k_v} \\ &= \beta_0 + \beta_1 \frac{\sum_{v\in S} \sum_{w\in S} A_{v,w}}{\sum_{v\in S} k_v} + \beta_2 \frac{\sum_{v\in S} \sum_{w\in S} \sum_{v_2=1}^{n} \frac{A_{v,v_2} A_{v_2,w}}{k_{v_2}}}{\sum_{v\in S} k_v}. \end{aligned} \tag{64}$$

For a small community with $C(S) \ll 1$, the community strength of a set S can be represented by equation (64), where the first term is simply a constant β_0 (for normalizing the community strength), the second term is equal to β_1(1−conductance), and the third term is related the clustering coefficient (in terms of the number of triangles) in the set S. In view of this, using a random walk with path length not greater than 2 to sample a network seems to be a good compromise between the viewpoint from conductance (Andersen et al., 2006; Andersen & Lang, 2006; Leskovec et al., 2008) and the viewpoint of clustering coefficient (Watts & Strogatz, 1998).

As β_0 is only a constant for normalizing the community strength, one can increase the community strength of a node by increasing β_0. Note that

$$Str(\{w\}) = C(\{w\}|\{w\}) - C(\{w\}) \geqslant \beta_0 - \frac{k_w}{2m}. \tag{65}$$

Thus, if we choose $\beta_0 \geqslant \frac{k_{\max}}{2m}$ (with $k_{\max} = \max_{v\in V_g} k_v$ being the maximum degree), then $Str(\{w\}) \geqslant 0$ for every node w in the sampled graph. In our experiments, we use the following coefficients: $\beta_2 = 0.25$, $\beta_0 = k_{max} / 2m$, and $\beta_1 = 1 - \beta_0 - \beta_2$. As such, every single node has a non-negative community strength and thus a community by itself.

5.2 *College football games*

Our first dataset is the American football games between Division IA colleges during regular season Fall 2000 (Girvan & Newman, 2002), where nodes represent teams and edges represent regular-season games between two teams. The teams are

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

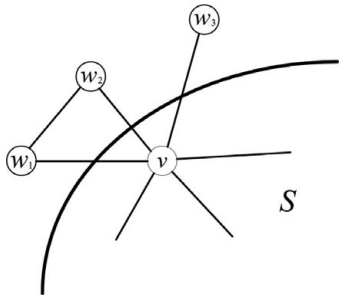


Fig. 1. An illustrating example with three nodes.



Fig. 2. The network of American football games between Division IA colleges during regular season Fall 2000. (Color online)

divided into 12 conferences (See Figure 2). As shown in Figure 2, the competitions (edges) in the same conference are more intense than those between two different conferences. We first compute the community strengths of these 12 conferences to see how strong these conferences (as communities) are. These results are shown in the second column of Table 1 and there are several conferences that have small community strengths, in particular conference 5 only has the community strength 0.04 (note that we have added β_0 in our algorithm to ensure that every node has a non-negative community strength).

To evaluate our algorithm, we take every team as a seed node and run our local community detection algorithm with the maximum size being set to the size of its conference. We then compute the average precision and recall for each conference in Table 1. As expected, conferences with larger community strengths are easier to be detected. In particular, for conferences 0,1,2,3,6,7,8, and 9, our local community detection algorithm achieves 100% precision and recall. All these conferences have community strengths larger than 0.52. On the other hand, our algorithm does not achieve high precision and recall for conference 5, which has the lowest community strength 0.04 among all the conferences. To further understand this, we plot the edges within conference 5 in Figure 3. There is only one edge between two teams in this conference! We also plot the edges within conferences 10 and 11 in Figures 4 and

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

Table 1. *Precision and recall for college football.*

Conference	Strength	Precision(%)	Recall(%)
0	0.63	100.00	100.00
1	0.54	100.00	100.00
2	0.57	100.00	100.00
3	0.60	100.00	100.00
4	0.46	82.00	82.00
5	0.04	24.00	24.00
6	0.59	100.00	100.00
7	0.52	100.00	100.00
8	0.60	100.00	100.00
9	0.61	100.00	100.00
10	0.24	51.02	51.02
11	0.43	66.00	66.00

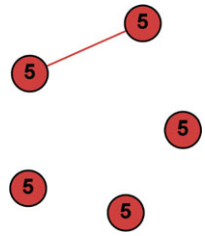


Fig. 3. Edges within conference 5 of American football games. (Color online)

5, respectively. These two conferences also have relative low community strengths and low precision and recall results. As shown in Figure 4, Conference 10 is formed by a single edge between a clique of four nodes and another clique of three nodes. As such, our algorithm detects this conference as two communities and thus achieves roughly 50% of precision and recall. As for conference 11, there is one team that did not play any games with any other teams in the same conference. As such, our algorithm is not able to detect that team for conference 11 and that results in poor precision and recall results.

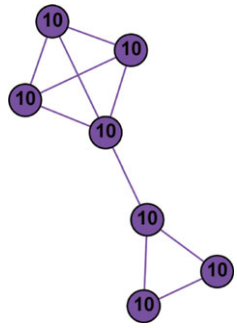


Fig. 4. Edges within conference 10 of American football games. (Color online)

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

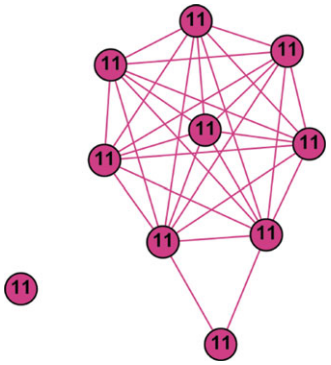


Fig. 5. Edges within conference 11 of American football games. (Color online)

Table 2. *Precision and recall for the Zachary karate club.*

Maximum relative centrality		
Precision(%)	Recall(%)	Number of nodes
100.00	100.00	5
100.00	94.44	15
100.00	77.78	2
91.67	68.75	11
8.33	5.56	1

5.3 Zachary karate club

Our second dataset is the Zachary karate club dataset (Zachary, 1977) that has been used for benchmarking various community detection algorithms in the literature. The karate club split into two groups due to a dispute in the club. Some members of the club left the original club and established another new club. In Figure 6, we reproduce the network of friendships in the karate club, where nodes represent members and edges represent friendships.

As described in our experiments for the American football games, we take every member as a seed node and run our local community detection algorithm with the maximum size being set to the size of his/her club. In Table 2, we show the precision and recall results for the Zachary karate club. Our algorithm achieves 100% precision for 22 nodes and 91.67% precision for another 11 nodes. The results for recall are not as good as those for precision. Only 5 nodes have 100% recall and another 15 nodes have 94.44% recall. There is one node, node 9 in Figure 6, that has poor precision and recall. This is because node 9 is at the boundary of these two clubs and there is only one link from node 9 to each club. As such, both node 2 and node 33 have the same relative centrality with respect to node 9. For our algorithm, we break the tie by choosing the node with a smaller index and thus node 9 chooses node 2 in the first step. Clearly, this then leads to the wrong clustering of node 9 to the other club.

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

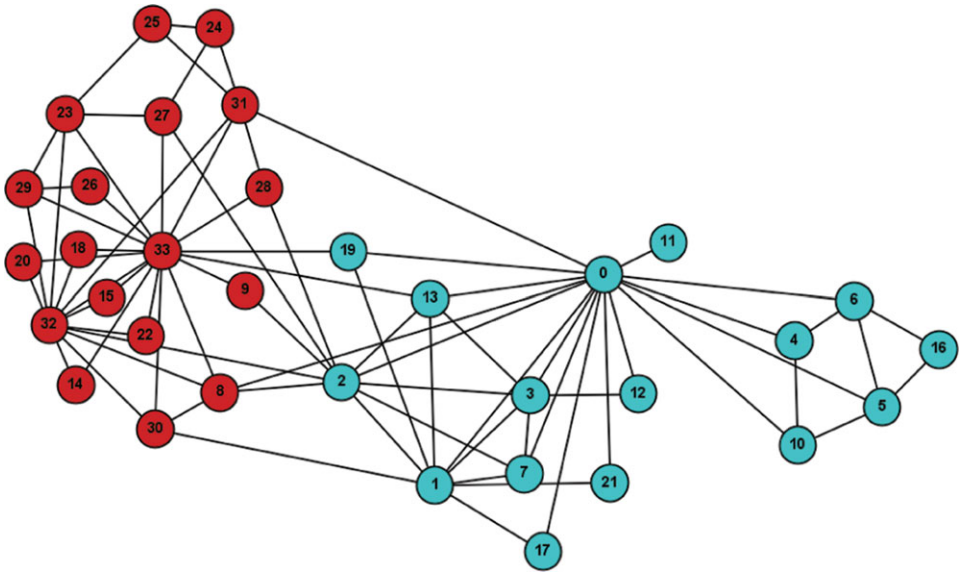


Fig. 6. The network of friendships in the Zachary karate club. (Color online)

5.4 LFR benchmark graphs

Both the dataset for the American football games and the dataset for the Zachary karate club are quite small. In order to test the scalability of the local community detection algorithm, we consider the LFR benchmark graphs (Lancichinetti et al., 2008) with $n = 5000$ nodes. In the LFR benchmark graphs, both the degree and the community size distributions are power laws, with exponents γ and β, respectively. As in Lancichinetti et al. (2008), the average degree of a node is denoted by $\langle k \rangle$. In addition to these three parameters, the mixing parameter μ is used for characterizing how the built-in communities in a graph are mixed. Specifically, for each node, a fraction $1 - \mu$ of its links are with the other nodes of its community and a fraction μ of its links are with the other nodes of the graph.

In each of our experiments, 20 realizations of LFR graphs are generated. For each built-in community in a realization, we then randomly choose a node as a seed node and run our local community detection algorithm with the maximum size being set to the size of the built-in community. By doing so, we have the same number of (overlapping) communities as that of the built-in communities. We then compute the normalized mutual information measure (NMI) by using a built-in function in GitHub (McDaid et al., 2011; Lancichinetti et al., 2009). In Figure 7, we show our experimental results for four pairs of the exponents $(\gamma, \beta) = (2, 1), (2, 2), (3, 1), (3, 2)$, and three values of the average degree $\langle k \rangle = 15, 20, 25$. Each curve (marked with RC) shows the variation of the NMI with respect to the mixing parameter μ.

To compare with other existing community detection methods, we also perform the same experiments by using the modularity optimization method (Blondel et al., 2008) (marked with FU in Figure 7) and the Potts model (Reichardt & Bornholdt, 2006b) (marked with PM in Figure 7) implemented in igraph (Csardi & Nepusz, 2006). We note the numerical results for the modularity optimization method and the Potts model are slightly different from those in Lancichinetti et al. (2008). This is

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

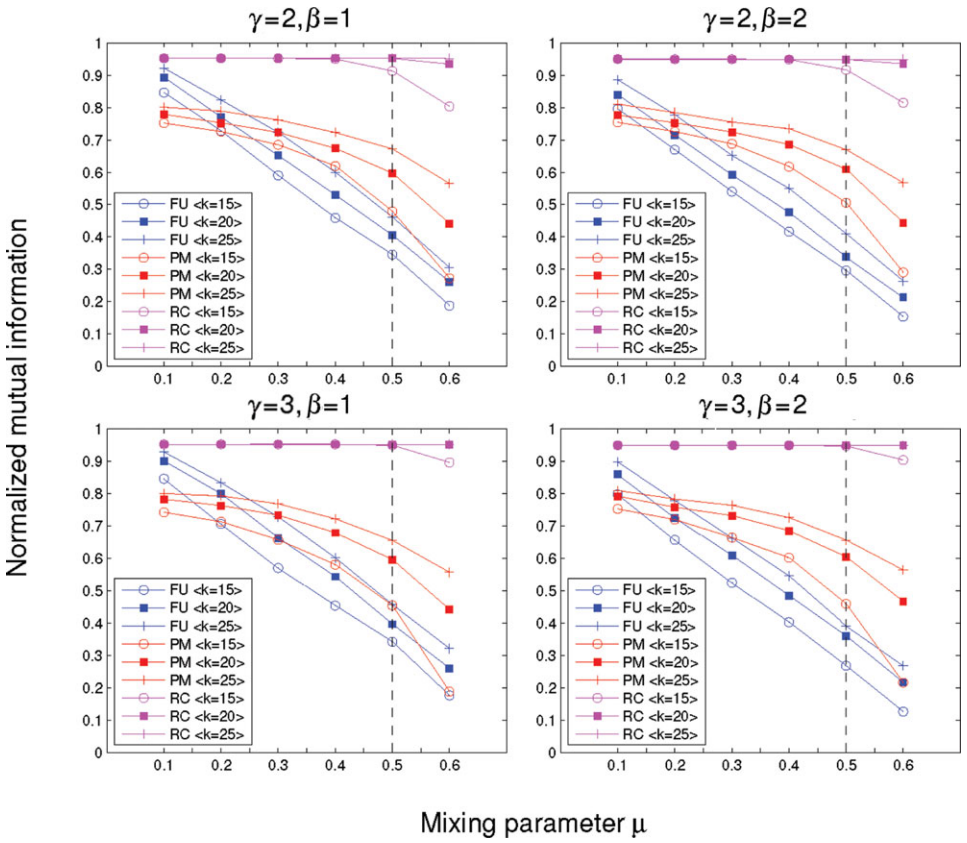


Fig. 7. Test of the local community detection algorithm on the LFR graphs. The number of nodes $n = 5000$. Each point corresponds to an average over 20 graph realizations. (Color online)

mainly due to the fact that the NMI used here allows overlapping communities and it is different from that in Lancichinetti et al. (2008). As can be seen from Figure 7, our local community detection algorithm gives better results in these experiments. To see the intuition behind this, let us consider the case with the mixing parameter $\mu = 0.6$ and $\langle k \rangle = 25$. In this case, the average number of links connected to the other nodes within the same community is $25 * (1 - 0.6) = 10$. On average, the other 15 links are distributed to a large number of communities. In our experiments, the average number of built-in communities is 154.5. As the LFR graphs are generated by using the configuration model, the probability that a node has more links to the other communities than its own community is quite small. As such, local search based on relative centrality can be very effective. On the other hand, if the average degree $\langle k \rangle$ is small, it is more likely for the local community detection algorithm to include a wrong node at the early stage of the algorithm and that might lead to a chain effect to include much more wrong nodes in the end. Finally, we note that these numerical results in terms of the NMI measures should not be regarded as a fair comparison with the modularity optimization method and the Potts model as these two methods are intrinsically different from our local community detection algorithm. For instance, in our experiments, we run our local community detection

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

algorithm with the size equal to the size of the built-in community. Such information is in general not available.

5.5 *Large real networks*

In this section, we perform our experiments by using the Amazon dataset and the DBLP dataset (Leskovec & Krevl, 2014). These two datasets are considerably larger than the American football games and the Zachary karate club that we used in the previous experiments. As indicated in Leskovec & Krevl (2014), the Amazon dataset was collected by crawling the Amazon website. In such a network, an undirected edge between product i and product j is added if product i is frequently co-purchased with product j. Each product category provided by Amazon defines each ground-truth community. Also, each connected component in a product category is regarded as a separate ground-truth community. In the Amazon dataset (Leskovec & Krevl, 2014), the number of nodes is 3,34,863, the number of edges is 9,25,872, and the number of ground-truth communities is 5,000.

On the other hand, the DBLP computer science bibliography provides a comprehensive list of research papers in computer science. A co-authorship network is constructed by adding an edge between two authors if they publish at least one paper together. Publication venue, e.g, journal or conference, defines an individual ground-truth community and authors who published to a certain journal or conference form a community. Also, each connected component in a group is regarded as a separate ground-truth community. In the DBLP dataset (Leskovec & Krevl, 2014), the number of nodes is 3,17,080, the number of edges is 10,49,866, and the number of ground-truth communities is 5,000.

In our experiments, we only test the largest 110 communities among the 5,000 ground truth communities in both datasets. For the parameters needed for the random walk with path length not greater than 2, we choose $\beta_0 = 0$, $\beta_2 = 0$, and $\beta_1 = 1$. In other words, we only use the information of the degree of a node in our local community detection algorithm. Instead of starting from a single node in our local community detection algorithm, we start from an initial seed set that contains more than one node. The initial seed set S_0 of each ground-truth community S is the set of nodes that contains the minimum number of nodes so as to satisfy $C(S_0|S) \geqslant \alpha C(S|S)$. Such a seed set can be viewed as a *core* of a set S and the parameter α determines the size of the core. In our experiment, the parameter α is 0.9. The main objective of these experiments is to see whether our local community detection algorithm is effective in detecting large ground truth communities by starting from a core of a community.

In Figure 8, we show the experimental results for the Amazon dataset in terms of community strength, community size and precision. The precision for each ground-truth community S in Figure 8 is the number of nodes that are correctly detected in the set $S \backslash S_0$ divided by the size of the set $S \backslash S_0$. As in the LFR benchmark graphs, we can see from Figure 8 that precision is positively correlated with community strength, especially for those communities with small community sizes. However, for communities with very large community sizes, the correlation between community strength and precision is not that clear.

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

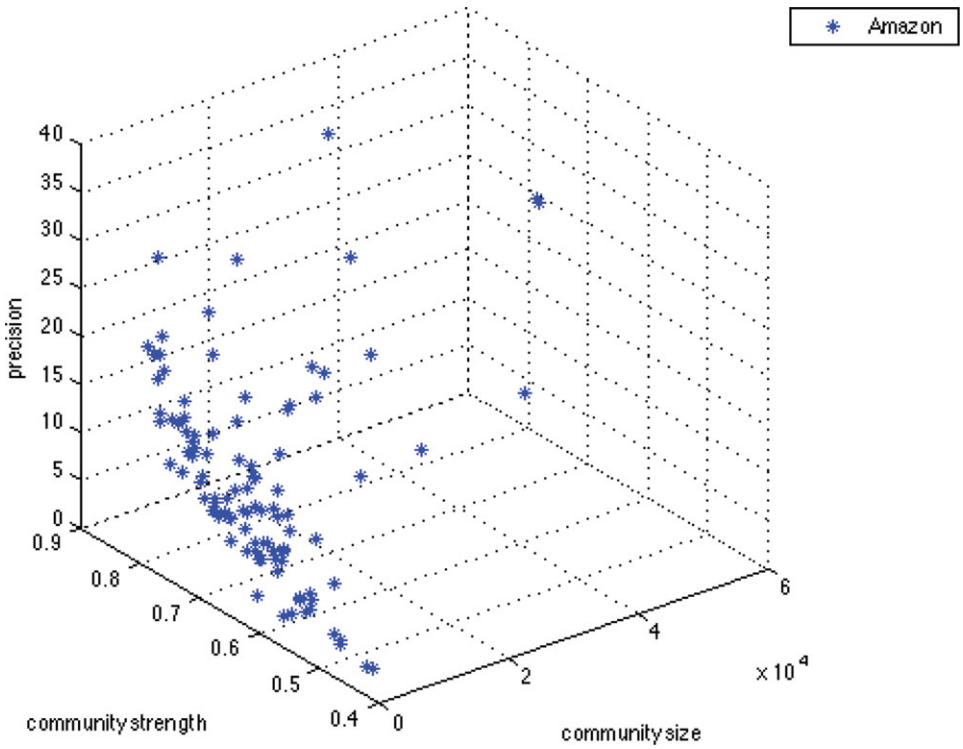


Fig. 8. Experimental results for the Amazon dataset. (Color online)

In Figure 9, we show the experimental results for the DBLP dataset. To our surprise, the correlation between community strength and precision is not clear even for communities with small community sizes. A careful examination of the ground truth communities in the DBLP dataset reveals that the Jaccard similarity index between the cores of a pair of two ground truth communities in the DBLP dataset could be high sometimes. In other words, the ground truth communities in the DBLP dataset are overlapping communities with significant amounts of overlaps in their cores. This might due to the fact that there are overlapping conference attendees in conferences on similar or related topics. In such a setting, greedy local community detection algorithms do not perform well. To see this, consider two sets S_1 and S_2 in Figure 10 that have significant overlaps in their cores. If we start from the core in S_1, we might end up with detecting the set S_2 by using a *greedy* local community detection algorithm. In such a scenario, the precision is quite low. To summarize, unlike the LFR benchmark graphs, the DBLP dataset have overlapping communities that have significant overlaps in their cores. As such, *greedy* local community detection algorithms like the one presented in this paper may not be effective in detecting a *specific* community that contains a core of that community. One tentative solution is not to use the greedy selection in (P5) of our local community detection algorithm. But this requires keeping track of more candidate nodes and thus increases computational complexity in the local community detection algorithm. On the other hand, we also note that the community detection algorithms in Blondel et al. (2008); Reichardt & Bornholdt (2006b) are designed for

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

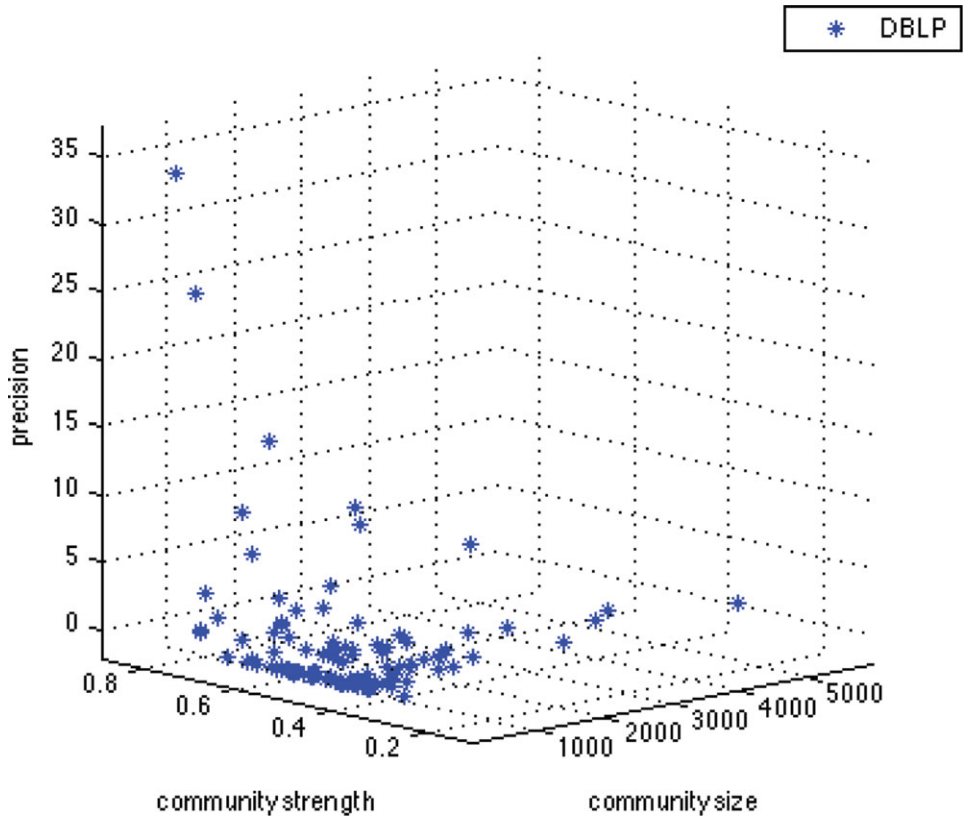


Fig. 9. Experimental results for the DBLP dataset. (Color online)

finding a *partition* of a network. As such, they are also not applicable for detecting overlapping communities.

6 Conclusion

In this paper, we make an attempt to lay a foundation for structure analysis of networks. For this, we introduced the concept of relative centrality for a sampled graph. We then developed the associated framework on top of that and illustrated its connections to many early works in the literature. In particular, the community strength of a set of nodes is defined as the difference between the relative centrality of the set with respect to itself and its centrality, and a community is then a set of nodes with a non-negative community strength. We showed there are several equivalent statements for a community in our framework. We also defined the modularity as the average community strength of the community to which a randomly selected nodes belong. Such a notion generalized the previous modularity in Newman & Girvan (2004) and stability in Lambiotte (2010); Delvenne et al. (2010).

For the local community detection problem, we developed agglomerative algorithms that guarantee the community strength of the detected local community. There are two key features of our local community detection algorithms: (i) local search: only nodes in the *neighboring set* need to be explored, and (ii) recursive update: the relative centralities can be efficiently updated by using recursive formulae.

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

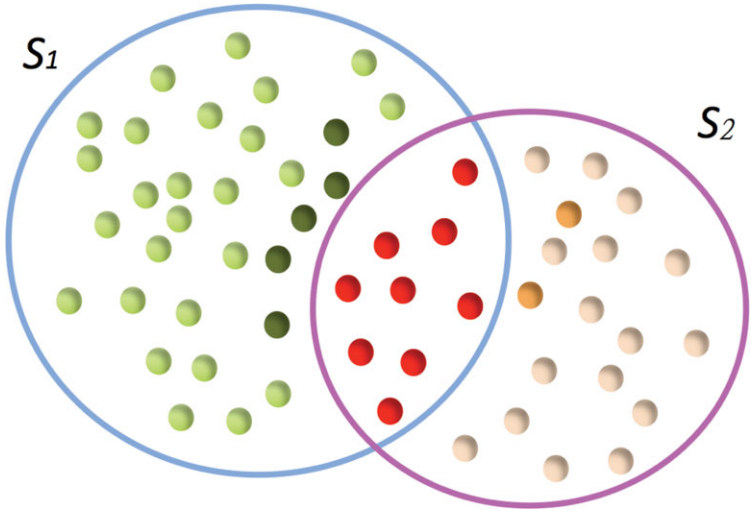


Fig. 10. Greedy local community detection algorithms do not perform well when the cores of the two sets S_1 and S_2 are similar. (Color online)

As such, they are as efficient as the algorithm in Clauset (2005) in terms of computational complexity. As our local community detection algorithm has the freedom to choose the "viewpoint" to sample a network, we propose sampling the graph with a random walk that has a path length not greater than 2. Such a sampling method allows us to look beyond the first neighbors so that we can take the clustering coefficient into account.

There are several possible extensions for our work:

i. Sampling with bivariate distributions that are not symmetric: here we assume that the bivariate distribution associated with a sampled graph is *symmetric*. Such a condition might be too strong for sampling directed networks. A possible extension is to consider bivariate distributions that have the same marginal distributions.

ii. Link prediction: as addressed in Liben-Nowell & Kleinberg (2003), similarity measures can also be used for link prediction. The concept and the framework developed for relative centrality here might be used for more complicated types of link prediction, e.g., Would some products be more preferable to one group of people than another group of people?

iii. Community detection: we note our algorithm is for the *local* community detection problem and it is different from the community detection problem. The local community detection problem is to find a community that contains a specific node(s). As such, one first needs to define what a community is (as we did in the paper). Also, one only needs the local information around that node and it is independent of the size of a network once some global statistics are known, e.g., the total number of nodes and the total number of edges. On the other hand, the community detection problem is to partition the whole network into communities. These are two different problems and it would not be a fair comparison between a local community detection algorithm and a community detection algorithm. For the community detection problem, it might have the scalability issue. However, for the local community detection problem, there is no scalability issue once the size

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

of the community that one would like to find is specified. One possible application of our local community detection algorithm is to develop an app on an on-line social network so that the app can automatically generate a list of friends to invite for a party. For such an app, there is no need to know the whole on-line social network. For the community detection problem, especially for detecting overlapping communities, we refer to the survey paper (Xie et al., 2013) for a comparative study and (Yang & Leskovec, 2013) for scalable detection of large overlapping communities.

Acknowledgment

This work was supported in part by Ministry of Science and Technology (MOST), Taiwan, under Grant Number 102-2221-E-007-006-MY3.

References

Andersen, R., Chung, F., & Lang, K. (2006). Local graph partitioning using pagerank vectors. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, 2006. focs'06*. IEEE pp. 475–486.

Andersen, R., & Lang, K. J. (2006). Communities from seed sets. In *Proceedings of the 15th International Conference on World Wide Web*. ACM, pp. 223–232.

Arenas, A., Fernandez, A., & Gomez, S. (2008). Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, **10**(5), 053039.

Bell, J. R. (2012). Relative centrality measures. Retrieved from http://www.westpoint.edu/nsc/SiteAssets/SitePages/Publications

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**(10), P10008.

Boyd, S., Ghosh, A., Prabhakar, B., & Shah, D. (2005). Gossip algorithms: Design, analysis and applications. In *Proceedings of the IEEE Infocom 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 3. IEEE, pp. 1653–1664.

Brandes, U., Robins, G., Mccranie, A., & Wasserman, S. (2013). What is network science? *Network science*, **1**(1), 1–15.

Chang, C.-S., Hsu, C.-Y., Cheng, J., & Lee, D.-S. (2011). A general probabilistic framework for detecting community structure in networks. In *Proceedings IEEE Infocom, 2011*. IEEE, pp. 730–738.

Clauset, A. (2005). Finding local community structure in networks. *Physical Review E*, **72**(2), 026132.

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

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *Interjournal*, **Complex Systems**, 1695.

Danon, L., Diaz-Guilera, A., Duch, J., & Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, **2005**(09), P09008.

Delvenne, J.-C., Yaliraki, S. N., & Barahona, M. (2010). Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences*, **107**(29), 12755–12760.

Dhillon, I. S., Guan, Y., & Kulis, B. (2004). Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of the 10th acm sigkdd International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 551–556.

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

Diaconis, P., & Stroock, D. (1991). Geometric bounds for eigenvalues of markov chains. *The Annals of Applied Probability*, **1**(1), 36–61.

Duch, J., & Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Physical Review E*, **72**(2), 027104.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, **486**(3), 75–174.

Fortunato, S., & Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, **104**(1), 36–41.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, **40**(1), 35–41.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, **1**(3), 215–239.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, **99**(12), 7821–7826.

Granell, C., Gomez, S., & Arenas, A. (2012). Hierarchical multiresolution method to overcome the resolution limit in complex networks. *International Journal of Bifurcation and Chaos*, **22**(07), 1250171.

Hu, Y., Chen, H., Zhang, P., Li, M., Di, Z., & Fan, Y. (2008). Comparative definition of community and corresponding identifying algorithm. *Physical Review E*, **78**(2), 026121.

Huang, X.-C., Cheng, J., Chou, H.-H., Cheng, C.-H., & Chen, H.-T. (2013). Detecting overlapping communities in networks based on a simple node bahavior model. In *Proceedings of IEEE Globecom 2013*, pp. 3120–3125.

Kamvar, D. S., Klein, D., & Manning, C. (2003). Spectral learning. *International Joint Conference of Artificial Intelligence*. Stanford InfoLab.

Karrer, B., & Newman, M. E. J. (2011). Stochastic block models and community structure in networks. *Physical Review E*, **83**(1), 016107.

Karrer, B., Levina, E., & Newman, M. E. J. (2008). Robustness of community structure in networks. *Physical Review E*, **77**(4), 046119.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, **18**(1), 39–43.

Kulis, B., Basu, S., Dhillon, I., & Mooney, R. (2009). Semi-supervised graph clustering: A kernel approach. *Machine Learning*, **74**(1), 1–22.

Lambiotte, R. (2010). Multi-scale modularity in complex networks. In *Proceedings of the 8th International Symposium on Modeling and Optimization in Mobile, ad hoc and Wireless Networks (wiopt)*. IEEE, pp. 546–553.

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E*, **80**(5), 056117.

Lancichinetti, A., & Fortunato, S. (2011). Limits of modularity maximization in community detection. *Physical Review E*, **84**(6), 066122.

Lancichinetti, A., Fortunato, S., & Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, **11**(3), 033015.

Lancichinetti, A., Fortunato, S., & Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical Review E*, **78**(4), 046110.

Leskovec, J., & Krevl, A. (2014) (June). SNAP Datasets: Stanford large network dataset collection. Retrieved from http://snap.stanford.edu/data.

Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. W. (2008). Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th International Conference on World Wide Web*. ACM, pp. 695–704.

Leskovec, J., Lang, K. J., & Mahoney, M. (2010). Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, pp. 631–640.

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

Liben-Nowell, D., & Kleinberg, J. (2003). The link prediction problem for social networks. In *Proceedings of the 12th International Conference on Information and Knowledge Management*. ACM, pp. 556–559.

Long, B., Zhang, Z. M., & Yu, P. S. (2007). A probabilistic framework for relational clustering. In *Proceedings of the 13th ACM sigkdd International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 470–479.

McDaid, A. F., Greene, D., & Hurley, N. (2011). Normalized mutual information to evaluate overlapping community finding algorithms. Oct. Retrieved from http://arxiv.org.

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., & Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, **328**(5980), 876–878.

Newman, M. (2009). *Networks: an introduction*. Oxford, England: OUP .

Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, **69**(6), 066133.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, **69**(2), 026113.

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**(7043), 814–818.

Porter, M. A., Onnela, J.-P., & Mucha, P. J. (2009). Communities in networks. *Notices of the AMS*, **56**(9), 1082–1097.

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 the United States of America*, **101**(9), 2658–2663.

Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, **76**(3), 036106.

Reichardt, J., & Bornholdt, S. (2006a). Statistical mechanics of community detection. *Physical Review E*, **74**(1), 016110.

Reichardt, J., & Bornholdt, S. (2006b). Statistical mechanics of community detection. *Physical Review E*, **74**(1), 016110.

Rosvall, M., & Bergstrom, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, **104**(18), 7327–7331.

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, **105**(4), 1118–1123.

Spielman, D. A., & Teng, S.-H. (2004). Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the 36th Annual acm Symposium on Theory of Computing*. ACM, pp. 81–90.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-worldnetworks. *Nature*, **393**(6684), 440–442.

Wu, F., & Huberman, B. A. (2004). Finding communities in linear time: A physics approach. *The European Physical Journal b-Bondensed Matter and Complex Systems*, **38**(2), 331–338.

Xie, J., Kelley, S., & Szymanski, B. K. (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (csur)*, **45**(4), 43.

Xu, B., Liang, Z., Jia, Y., Zhou, B., & Han, Y. (2012). Local community detection using seeds expansion. In *Proceedings of the 2nd International Conference on Cloud and Green Computing (cgc)*. IEEE, pp. 557–562.

Yang, J., & Leskovec, J. (2012). Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM sigkdd Workshop on Mining Data Semantics*. ACM, p. 3.

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

Yang, J., & Leskovec, J. (2013). Overlapping community detection at scale: A nonnegative matrix factorization approach. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*. ACM, pp. 587–596.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, **33**(4), 452–473.

Appendix A

In this section, we prove Theorem 3.1.

(i) $\Rightarrow$ (ii): Note from Proposition 2.1 (i) and (ii) that $C(S|S)+C(S^c|S) = C(V_g|S) = 1$ and $C(S) + C(S^c) = C(V_g) = 1$. It then follows from the reciprocal property in Proposition 2.1(iv) that

$$\begin{aligned} &C(S^c)(C(S|S) - C(S|S^c)) \\ &\quad = C(S^c)C(S|S) - C(S^c)C(S|S^c) \\ &\quad = (1 - C(S))C(S|S) - C(S)C(S^c|S) \\ &\quad = (1 - C(S))C(S|S) - C(S)(1 - C(S|S)) \\ &\quad = C(S|S) - C(S) = Str(S) \geqslant 0. \end{aligned}$$

As we assume that $0 < C(S) < 1$, we also have $0 < C(S^c) < 1$. Thus,

$$C(S|S) - C(S|S^c) \geqslant 0.$$

(ii) $\Rightarrow$ (iii): Since we assume that $C(S|S) \geqslant C(S|S^c)$, we have from $C(S|S) + C(S^c|S) = C(V_g|S) = 1$ that

$$1 = C(S|S) + C(S^c|S) \geqslant C(S|S^c) + C(S^c|S).$$

Multiplying both sides by $C(S^c)$ yields

$$C(S^c) \geqslant C(S^c)C(S|S^c) + C(S^c)C(S^c|S).$$

From the reciprocal property in Proposition 2.1(iv) and $C(S)+C(S^c) = C(V_g) = 1$, it follows that

$$\begin{aligned} C(S^c) &\geqslant C(S)C(S^c|S) + C(S^c)C(S^c|S) \\ &= (C(S) + C(S^c))C(S^c|S) \\ &= C(S^c|S). \end{aligned}$$

(iii) $\Rightarrow$ (iv): Note from the reciprocal property in Proposition 2.1(iv) that

$$C(S)C(S^c|S) = C(S^c)C(S|S^c). \tag{66}$$

It then follows from $C(S^c|S) \leqslant C(S^c)$ that $C(S|S^c) \leqslant C(S)$.

(iv) $\Rightarrow$ (i): Since we assume that $C(S|S^c) \leqslant C(S)$, it follows from equation (66) that $C(S^c|S) \leqslant C(S^c)$. In conjunction with $C(S|S) + C(S^c|S) = C(V_g|S) = 1$ and $C(S) + C(S^c) = C(V_g) = 1$, we have

$$C(S|S) - C(S) = C(S^c) - C(S^c|S) \geqslant 0.$$

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

Appendix B

In this section, we prove Lemma 4.1.

i. Since S_1 and S_2 are disjoint, we have from the reciprocity property and the additivity property in Proposition 2.1 (iv) and (ii) that

$$\begin{aligned} & C(S_1 \cup S_2) \cdot C(S_3|S_1 \cup S_2) \\ &= C(S_3) \cdot C(S_1 \cup S_2|S_3) \\ &= C(S_3) \cdot C(S_1|S_3) + C(S_3) \cdot C(S_2|S_3) \\ &= C(S_1) \cdot C(S_3|S_1) + C(S_2) \cdot C(S_3|S_2). \end{aligned}$$

From the additivity property in Proposition 2.1 (ii) for two disjoint sets, we also have

$$C(S_1 \cup S_2) = C(S_1) + C(S_2).$$

Thus,

$$C(S_3|S_1 \cup S_2) = \frac{C(S_1) \cdot C(S_3|S_1) + C(S_2) \cdot C(S_3|S_2)}{C(S_1) + C(S_2)}.$$

ii. From the definition of community strength in equation (38), we have that

$$Str(S_1 \cup S_2) = C(S_1 \cup S_2|S_1 \cup S_2) - C(S_1 \cup S_2) \tag{67}$$

$$Str(S_1) = C(S_1|S_1) - C(S_1) \tag{68}$$

and

$$Str(S_2) = C(S_2|S_2) - C(S_2). \tag{69}$$

Using (i) of this lemma and the additivity property in Proposition 2.1 (ii) for two disjoint sets yields

$$\begin{aligned} & C(S_1 \cup S_2|S_1 \cup S_2) \\ &= \frac{C(S_1) \cdot C(S_1 \cup S_2|S_1) + C(S_2) \cdot C(S_1 \cup S_2|S_2)}{C(S_1) + C(S_2)} \\ &= \frac{C(S_1) \cdot C(S_1|S_1) + C(S_1) \cdot C(S_2|S_1)}{C(S_1) + C(S_2)} + \frac{C(S_2) \cdot C(S_1|S_2) + C(S_2) \cdot C(S_2|S_2)}{C(S_1) + C(S_2)}. \end{aligned} \tag{70}$$

In view of the reciprocity property in Proposition 2.1 (iv),

$$C(S_1) \cdot C(S_2|S_1) = C(S_2) \cdot C(S_1|S_2). \tag{71}$$

Thus,

$$\begin{aligned} & C(S_1 \cup S_2|S_1 \cup S_2) \\ &= \frac{C(S_1) \cdot C(S_1|S_1) + 2C(S_1) \cdot C(S_2|S_1)}{C(S_1) + C(S_2)} + \frac{C(S_2) \cdot C(S_2|S_2)}{C(S_1) + C(S_2)}. \end{aligned} \tag{72}$$

From the additivity property in Proposition 2.1 (ii) for two disjoint sets, we also have

$$C(S_1 \cup S_2) = C(S_1) + C(S_2). \tag{73}$$

Using equations (72), (73), (68), and (69) in equation (67) yields the result in equation (51).

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press

If, furthermore, S_1 and S_2 are positively correlated, then it follows from equations (49) and (51) that

$$\begin{aligned} &Str(S_1 \cup S_2) \\ &\geqslant \frac{C(S_1) \cdot Str(S_1) + C(S_2) \cdot Str(S_2)}{C(S_1) + C(S_2)} \\ &\geqslant \min[Str(S_1), Str(S_2)]. \end{aligned} \tag{74}$$

Appendix C

In this section, we prove Proposition 4.1,

i. If w is not in $Nei(S) \cup S$, we have $C(\{w\}|S) = 0$. Since $C(\{w\}) > 0$, it follows that $C(\{w\}|S) - C(\{w\}) < 0$. Thus, w is not positively correlated to S.

ii. Since w^* is not in S, we know that $\{w^*\}$ and S are disjoint. From Lemma 4.1 (i), it follows that

$$C(\{w\}|S \cup \{w^*\}) = \frac{C(S) \cdot C(\{w\}|S) + C(\{w^*\}) \cdot C(\{w\}|\{w^*\})}{C(S) + C(\{w^*\})}. \tag{75}$$

Since $C(S) > 0$ and $C(\{w\}) > 0$, $C(\{w\}|S \cup \{w^*\}) > 0$ if and only if either $C(\{w\}|S) > 0$ or $C(\{w\}|\{w^*\}) > 0$. For any node w not in $S \cup \{w^*\}$, we then have $C(\{w\}|S \cup \{w^*\}) > 0$ if and only if w is in $Nei(S)$ or w is in $Nei(\{w^*\})$. As such,

$$Nei(S \cup \{w^*\}) = \Big(Nei(S) \cup Nei(\{w^*\})\Big) \backslash (S \cup \{w^*\}). \tag{76}$$

Write

$$Nei(S) \cup Nei(\{w^*\}) = Nei(S) \cup \Big(Nei(\{w^*\}) \backslash Nei(S)\Big).$$

From equation (76), we then have

$$\begin{aligned} &Nei(S \cup \{w^*\}) \\ &= \Big(Nei(S) \backslash (S \cup \{w^*\})\Big) \cup \Big((Nei(\{w^*\}) \backslash Nei(S)) \backslash (S \cup \{w^*\})\Big) \\ &= \Big(Nei(S) \backslash \{w^*\}\Big) \cup \Big(Nei(\{w^*\}) \backslash (S \cup Nei(S))\Big) \end{aligned}$$

where we use the fact that a node in $Nei(S)$ (resp. $Nei(\{w^*\})$ is not in S (resp. $\{w^*\}$) in the last identity.

iii. If a node w is in $Nei(\{w^*\}) \backslash (S \cup Nei(S))$, then w is not in $Nei(S)$ and S. Thus, $C(\{w\}|S) = 0$. The result in equation (55) then follows directly from equation (75).

https://doi.org/10.1017/nws.2015.23 Published online by Cambridge University Press