Social Networks 40 (2015) 154–162

Contents lists available at ScienceDirect

Social Networks

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

On efficient use of entropy centrality for social network analysis and community detection

Alexander G. Nikolaev*, Raihan Razib, Ashwin Kucheriya

Department of Industrial and Systems Engineering, 438 Bell Hall, State University of New York at Buffalo, Buffalo, NY 14260, United States

ARTICLE INFO

Keywords:
Social network modeling
Centrality
Entropy
Community detection
Clustering

ABSTRACT

This paper motivates and interprets entropy centrality, the measure understood as the entropy of flow destination in a network. The paper defines a variation of this measure based on a discrete, random Markovian transfer process and showcases its increased utility over the originally introduced path-based network entropy centrality. The re-defined entropy centrality allows for varying locality in centrality analyses, thereby distinguishing locally central and globally central network nodes. It also leads to a flexible and efficient iterative community detection method. Computational experiments for clustering problems with known ground truth showcase the effectiveness of the presented approach.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Despite the abundance of existing methods for measuring centrality in social networks, new research challenges and opportunities continue to emerge. In application to large network datasets, computational efficiency of evaluation becomes a major indicator of utility of centrality measures. Even more importantly, the typically reliable path-based measures lose sensitivity when the number of paths contributing to their formulae grows too large, making the evaluation of node centrality with respect to nearby neighbors (as opposed to the whole network) particularly difficult. In searching for answers to new challenges, it is desirable to design centrality measures with solid grounding in theory, while not compromising interpretability sought by social science practitioners.

This paper develops a centrality measure whose computation for a given node does not require dyad-based path enumeration. Instead, the presented measure relies on an absorbing Markovian process evolving over finite time – this allows for matrix multiplication-based computation of centrality. Depending on the absorption rate and evolution time, the presented measure enables centrality analysis at varying localities around a node of interest, thereby distinguishing locally central and globally central network nodes. The measure offers an information theory-based approach to measuring centrality, and takes a particular, previously unoccupied spot in the typology of flow-based centrality metrics.

Different measures of centrality capture different aspects of what it means for a node to be "central" to the network. In his seminal paper, Freeman (1979) argued that node degree centrality, the number of direct links incident to a node, indexes the node's activity; node betweenness centrality, based on the position of a node with respect to the all-pair shortest paths in a network, exhibits the node's potential for network control; and closeness centrality, the sum of geodesic distances from a node to all the other nodes, reflects its communication independence or efficiency. Borgatti (2005) conceptualized a typology of centrality measures based on the ways that traffic flows through the network. Two characteristics – the route the traffic follows (geodesics, paths, trails, or walks) and the method of propagation (parallel duplication, serial duplication, or transfer) – define the two-dimensional typology. Each measure of centrality makes assumptions about the importance of the various types of traffic flow, and hence, each measure of centrality can be assessed by where it falls in the typology. For example, betweenness centrality is perfect for networks featuring flows along geodesics. A node with high betweenness centrality is essentially a traffic checkpoint that can shut down the flow. At the same time, betweenness is an inappropriate measure in networks where flow is not constrained to follow geodesics. Non-geodesic paths avoid the checkpoints altogether, making an alternative measure essential. Over the years, researchers have proposed a number of different centrality measures, including eigenvector centrality (Bonacich, 1972), information centrality (Stephenson and Zelen, 1989), subgraph centrality (Estrada and Rodriguez-Velazquez, 2005), alpha centrality (Bonacich and Lloyd, 2001), etc. However, their meaning

* Correspondence author. Tel.: +1 716 645 4710.
E-mail address: anikolae@buffalo.edu (A.G. Nikolaev).

http://dx.doi.org/10.1016/j.socnet.2014.10.002
0378-8733/© 2014 Elsevier B.V. All rights reserved.

with respect to Borgatti's typology have not always been clearly defined or analyzed.

Tutzauer (2007) began to address this issue and proposed a centrality measure for networks characterized by path-based transfer flows. The path-based transfer model assumes that an object travels from a particular node (the one whose centrality is being evaluated) to a destination (the node itself or one of its neighbors) along a random path. More specifically, a path is sequentially built: if the flow originating node is randomly selected to be the next in the sequence, then the flow is over before it begins; otherwise, the object is randomly passed to one of the original node's immediate neighbors. Given that the object has arrived to the new node, the next transfer step destination is then randomly chosen from among its neighbors (including the current node, but not including any of the previously visited nodes), and again the flow either stops (if the current node in the sequence is selected) or continues on in the same fashion (if a different node is selected). For the described transfer model, the centrality of a given node can be defined as the entropy of the transfer's final destination. In other words, it can be expressed via the probabilities of transfer paths from the node to each of the other nodes. Despite the fact that the motivation for this entropy-based measure is intuitively and technically clear, the research community has been slow to adopt it for application purposes, largely due to the need for exhaustive path enumeration in evaluating the defined centrality.

This paper develops the idea of Tutzauer (2007), and presents a new, high-utility entropy centrality measure based on a discrete Markovian transfer process. In the presented model, a transferred object randomly walks through a network; then, the resulting measure – the walk destination entropy – can be efficiently computed, which opens new ways for insightful, computationally efficient analyses of networks. The structure of the paper is as follows. Section 2 introduces essential notation and the fundamentals of path-transfer flow process, builds a Markov model for the study of this process, presents an expression for the entropy centrality measure, and offers an illustrative computational example. Section 3 uses entropy centrality to design an algorithm for community detection in networks, and reports computational results with the algorithm applied to clustering problems with known ground truth. Section 4 offers discussion and concluding remarks.

2. Model description

2.1. Mathematical preliminaries

The mathematical representation of a network is a directed or undirected graph $G=(V, E)$, where $V=\{1, 2, \ldots, N\}$ is a finite, nonempty set of nodes (vertices), and E is a relation (a tie configuration) on V. The elements of E are called edges. The edge $(i, j) \in E$ is incident with the vertices i and j, and i and j are incident with the edge $(i, j) \in E$. Moreover, $(i, j) \in E$ is a link if $i \neq j$ and a loop if $i=j$. The incidence matrix of G has elements (b_{ij}), $i=1, 2, \ldots, N$, $j=1, 2, \ldots, N$ such that $b_{ij}=1$ if nodes i and j in the network are connected with an edge and 0 otherwise.

2.2. Centrality and entropy connection

To motivate the connection between the centrality of a given node and the concept of entropy, consider a network of friends transferring an object among themselves. The more central the original node is, the more difficult it is to predict the object's final destination. If the node is central, the object has a greater probability of traveling far in multiple potential directions. In contrast, a less central node has a more limited choice of immediate transfer options and the process is more likely to stop (be absorbed) before the number of transfer options increases, which makes its destination more predictable.

This idea can be more easily understood if one considers an extreme example of a network of one extrovert person and many introverts. An introvert is a node in the network with no or very few incident links, while an extrovert is a node adjacent to many nodes in the network. Assume that, according to a random rule, an object transfer process can terminate after the object is passed from one node to another, i.e., the object will eventually be absorbed by some node, termed destination node. In the case of high absorption probabilities, if the object transfer process originates from the extrovert (following the transfer process described above), the probability that it ends up at any given node is close to $1/N$. In contrast, if the transfer process originates from the introvert, then the flow first needs to reach the hub to go beyond it, limiting the likelihood that "far-away" nodes are reached at all.

The level of uncertainty of object destination, as a function of its origin, can be captured as destination entropy. The concept of entropy was first introduced in physics, and later, developed in information and communication sciences; entropy enjoys distinct and intuitive interpretations in multiple applied domains. In adopting it for the use in social network analysis, one avoids having to assess a node's position with respect to paths connecting all node pairs, and instead, focuses on the node's potential to diversify flow propagation.

2.3. Path transfer and random walk flows as foundations for entropy centrality computation

In assessing the value of node position using network flow, researchers have historically focused on *paths* as channels that flow may follow. Entropy centrality does not explicitly measure the ability of a node to interfere with path-based exchanges between other nodes; instead, it views a node of interest as flow originator.

The treatment of paths and flow types, relevant to the concept of entropy centrality, deserves a more in-depth discussion. This paper's contribution to centrality theory is akin to that of Newman (2005), who first proposed to use walks, instead of only shortest paths, for betweenness measurement. In entropy centrality calculation, the idea of analyzing random walks is further developed, by allowing walks to be randomly interrupted; the longer a given planned object route, i.e., the more exchanges (transfers) it requires, the less likely it is to be completed. To further illustrate this point, a review and discussion of path-transfer flows is in order.

Examples of path-transfer flows are aplenty among trading and smuggling networks (Tutzauer, 2007), especially when the traded or smuggled commodity is discrete such as the case of exotic animals, nuclear weapons material and parts, fossils, artworks and antiquities, and even trafficking humans. For a more peaceful example, consider a group of people linked by friendship ties, with one of them having a specific object. To model a path-transfer process, think of the object being passed from one person to another. The flow (i.e., object transfer) originates at a particular person in the group (i.e., a node in the graph). If that person does not pass the object to any one of their immediate friends, the flow is over before it begins; otherwise, the object flows (i.e., is transferred) to a randomly selected person. The next person then chooses whether to pass the object to their immediate friends, and again the flow either stops or continues. The object thus traverses a path in the network, traveling along the links, stopping when the process is absorbed at some node or if the object's trajectory completes a loop. According to the original model formulation, each of the eligible neighbors is assumed to be selected with equal likelihood, although this assumption can be relaxed without loss of generality. The main restriction in the path-transfer process is that the object cannot be passed to the nodes it has already visited.

This paper relaxes the restriction for flow to follow paths in the entropy centrality definition. Instead, it develops a model based on a special case of random walk, where each node has a positive probability of absorbing the flow for good (Newman, 2005; Noh and Rieger, 2004). The motivation for this alternative definition of the entropy centrality computation mechanism is two-fold. First, the path-based centrality is extremely difficult to use in practice. The necessity for complete path enumeration in its computation makes the original measure (Tutzauer, 2007) not suitable for the analysis of well-connected networks containing over ten nodes. On the contrary, the relevant transfer and absorption probabilities for random walks can be easily calculated using matrix-analytic methods. Second, note an important nuance in the entropy centrality concept that can be utilized with its definition based on random walks. Entropy centrality is calculated using a serial transfer model, however, because multiple transfer destination probabilities enter the entropy expression simultaneously, it may be more conducive to analyzing serial duplication processes. An iterative, step-by-step analysis of a random walk originating from a given node would inform one of the temporal (periodic) dynamics of the flow destination entropy, i.e., indicate how fast (in how many periods) the network can be informed/conquered if the spread of influence is initiated from the given node. Consider modeling a community becoming engaged into discussing a pertinent topic/issue picked up by one of its members from a news outlet. All the sequences in which people can converse come into play, and some conversations can occur simultaneously, as long as multiple community members are informed of the topic/issue by their neighbor(s). Thoughtful, as opposed to gossip-generating, conversations between people are rarely broadcast, they take place sequentially: the same news can be discussed by the same two individuals multiple times (think more about mulling over a political situation, rather than sharing a news of a rock-star making appearance at a night club).

In summary, walk-based entropy centrality can be most useful for identifying influential community members with respect to serial duplication process. This observation defines the measure's place in Borgatti's typology, reaffirming the motivation for introducing it.

2.4. Markov model and entropy centrality

Consider a connected network represented by graph $G=(V, E)$, with V being a set of N nodes indexed 1 through N, and with E being a relation on V. Refer to Fig. 1 for an illustrative example of a small network with $N=6$ nodes and $|E|=8$ edges. In a random walk based flow process, the immediate destination of an object transferred from an object-holder depends only on the current object position, and not on the sequence of nodes that the object visited prior to the current state, therefore, its position over time (in time periods) can be modeled as a Markovian process, or a Markov chain. For example, an object being transferred over the network in Fig. 1 could move from node 1 to node 4, and then in the next period, back from node 4 to node 1. It is also assumed that each node has an option to hold the object to itself in any given period even though it is connected to other nodes, thus taking a pause in communication. Additionally, each node can stop the flow for good, with the probability of such an event referred to as absorption probability a_v, $v \in V$. Fig. 2 depicts the node absorption probabilities fixed at $a=0.2$, which implies that in a single period node 1 can transfer the object to three nodes (i.e., self, node 2 or node 4) with the same probability of $(1-0.2)/3=0.27$. Fig. 2 also adds auxiliary nodes to the original network: labeled with apostrophes, these nodes represent absorbing states of the Markov chain. Note that in order to avoid cluttering in Fig. 2, the loop transitions are not depicted on it. Consider a stochastic process with the state diagram as given by Figs. 1 and 2 combined (including both the loops and absorbing

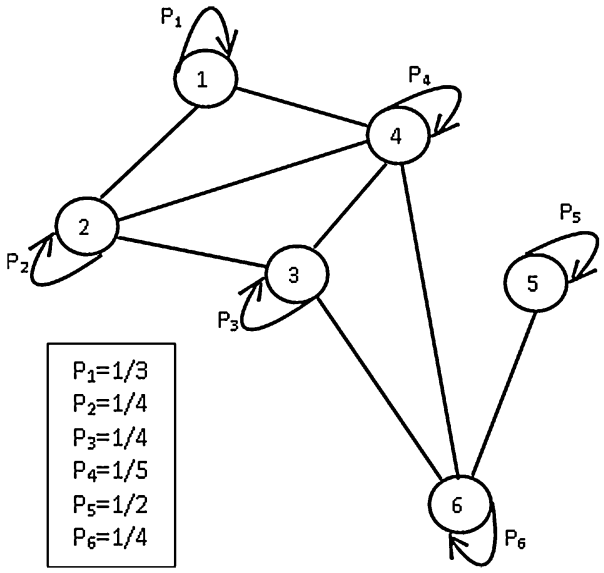

Fig. 1. A schematic representation of an example network depicting the object transfer (transition) diagram.

states). This process is a Markov chain with transition probability matrix denoted by P, with elements p_{ij}, $i \in \{1, 2, \ldots, N, 1', 2', \ldots, N'\}$, $j \in \{1, 2, \ldots, N, 1', 2', \ldots, N'\}$, as given in Table 1. The measure of centrality for node $i=1, 2, \ldots, N$, quantified by the entropy of the object destination, given that the transferred object originates from node i and experiences t transitions is defined as

$$H_i^t = -\sum_{j=1}^{N} (p_{ij}^{(t)} + p_{ij'}^{(t)}) \log(p_{ij}^{(t)} + p_{ij'}^{(t)}). \quad (1)$$

If the base of the logarithm in formula (1) is chosen to be 2, then the entropy centrality is measured in bits; meanwhile, the results in the subsequent sections of this paper are reported using the more conventional natural logarithm. The expression in (1) involves terms of the form $(p_{ij}^{(t)} + p_{ij'}^{(t)})$ – one such term gives the probability that the object originates at node i, and as t time periods elapse, finds itself in possession of node j. The closer these probabilities are for nodes $j \in \{1, 2, \ldots, N\}$, the more difficult it is to

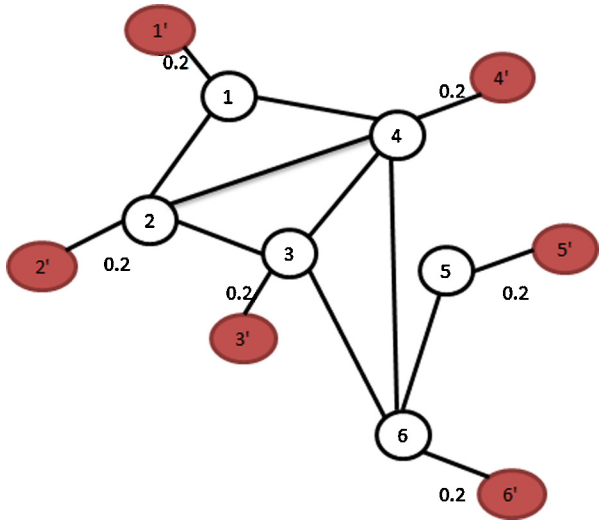

Fig. 2. An expanded state diagram for a Markovian transfer process, with auxiliary nodes for absorbing states.

Table 1
Transition probability matrix for the Markovian transfer process.

Node	1	2	3	4	5	6	1′	2′	3′	4′	5′	6′
1	0.267	0.267	0	0.267	0	0	0.2	0	0	0	0	0
2	0.2	0.2	0.2	0.2	0	0	0	0.2	0	0	0	0
3	0	0.2	0.2	0.2	0	0.2	0	0	0.2	0	0	0
4	0.16	0.16	0.16	0.16	0	0.16	0	0	0	0.2	0	0
5	0	0	0	0	0.4	0.4	0	0	0	0	0.2	0
6	0	0	0.2	0.2	0.2	0.2	0	0	0	0	0	0.2
1′	0	0	0	0	0	0	1	0	0	0	0	0
2′	0	0	0	0	0	0	0	1	0	0	0	0
3′	0	0	0	0	0	0	0	0	1	0	0	0
4′	0	0	0	0	0	0	0	0	0	1	0	0
5′	0	0	0	0	0	0	0	0	0	0	1	0
6′	0	0	0	0	0	0	0	0	0	0	0	1

predict/guess the object's position (at time t), and the larger is the entropy.

Note that the number of transitions t, which can be fixed at any integer, defines the desired locality of centrality analysis: it is thus termed *transfer locality*. In particular, a node in a network may have a high relative centrality for small t but low relative centrality for large t. Also, as t increases, the centrality measure for any node approaches a constant, depending on how fast the process is expected to be absorbed.

2.5. The effect of transfer locality adjustment

Given a locality value t, one evaluates a node's centrality with respect to a part of the network that is likely to be reached from the node by a transfer process in a limited time, i.e., in t steps of a random walk. In other words, by adjusting transfer locality, one "magnifies" the local neighborhood surrounding the node, thus reducing the impact of "far away", hard-to-reach nodes on the resulting entropy centrality value. When t is large, entropy centrality describes nodes' network positions on a global (whole network) scale.

In order to understand the implications of varying transfer locality in centrality analyses, consider a social network of Zackary's karate club. In a classical study, 34 members of a karate club were observed over a 2-year period. A network of friendships between the club members was constructed using a variety of measures to estimate the strength of ties. An unweighted version of the club network is given in Fig. 3; the following analysis focuses on the six nodes labeled 1, 5, 12, 29, 33 and 34 – these appear in bold circles in the figure. Fig. 4 reports entropy centrality values for the six nodes, with varied levels of t. With the increasing transfer locality, each node's centrality value monotonically increases, implying that, given more time, the node can reach more peers (remember, the node for which the centrality is computed is viewed as flow originator). Importantly, observe that the rates of entropy increase as a function of t are different for different nodes. As such, nodes 5 and 29 see their centrality values dramatically increase with the growing t, indicating that such nodes can become influential only if the transfer process they originate does not die early. Meanwhile, nodes 1, 33 and 34 located at "the heart" of well-connected clusters (small or large) do not see their centralities grow by much. Interestingly, node 29 has low centrality in its local neighborhood and high centrality with respect to the whole network, surpassing locally central nodes 5 and 33. The sensitivity of entropy centrality to a node's position with respect to network clusters leads to the idea of fixing the transfer locality value in such a way that clusters can be identified in any network.

3. Community structure detection

This section describes how entropy centrality can be used to reveal community structure in networks. The presented idea of a community detection algorithm is inspired by the algorithm proposed by Girvan and Newman (2002), which iteratively removes high-betweenness edges in an hierarchical clustering procedure. The algorithm proposed in this paper also removes one edge at a time and re-computes the corresponding transition and absorption probabilities for each node.

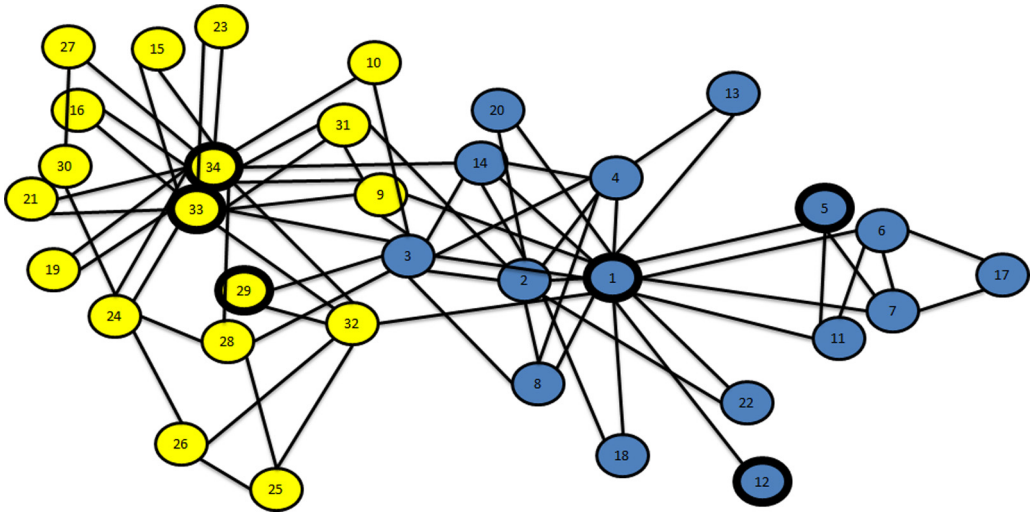

Fig. 3. Zackary's karate club social network.

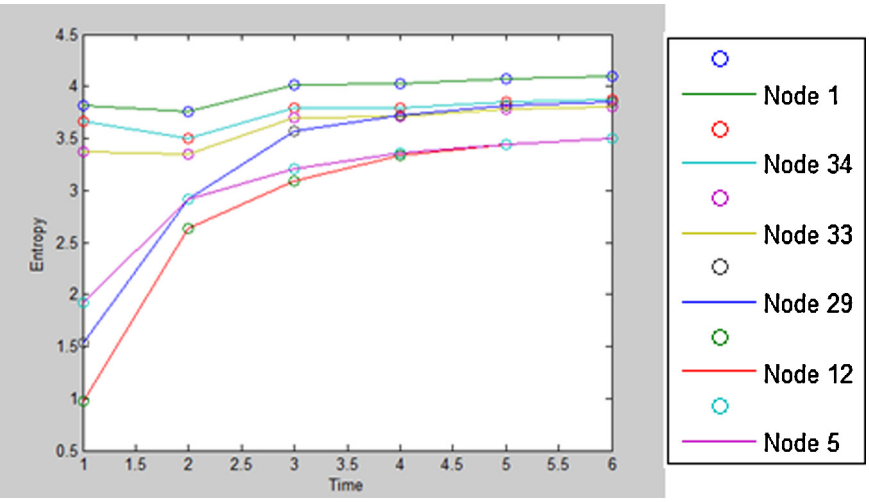

Fig. 4. Entropy centrality vs transfer locality plot for karate club problem.

1: **INPUT:** Number of nodes in the network N; transition probability matrix P for the Markov Chain with auxiliary nodes; transfer locality t; number of algorithm iterations K.
2: For $k = 1$ to K
3: For $i = 1$ to N and $j = 1$ to N, $i \neq j$
4: Remove the link between nodes i and j, if it exists
5: Revise the probabilities for transitions from nodes i and j
6: Compute the average entropy over all the nodes using (1)
7: Remember / Update the link for which the entropy decrease is maximum
8: End
9: Remove the link for which the entropy decrease is maximum
10: End
11: Sort to identify the obtained clusters
12: **OUTPUT:** The clusters.

Given an undirected graph and a fixed value of absorption probability for all the nodes, the transition probability matrix P for a Markov chain with auxiliary states is created first, as explained in Section 2. The desired centrality locality t is chosen next; during the experimentation, it was empirically discovered that locality values close to the diameter of a given network, and the absorption probability values in the range [0.1, 0.2] are convenient choices for successful global community detection. The algorithm proceeds by identifying and removing network edges such that the average entropy centrality over all the nodes is reduced the most (the algorithm pseudocode is given). Empirical investigations with the designed community detection algorithm to discover clusters in networks with known ground truth are reported next.

3.1. The Zachary's karate club network

Returning to the Zackary's karate club experiment, recall the part of the club's story that made it famous in the social network analysis circles: during the 2-year observational study, a split occurred between the club members. A disagreement, which developed between the administrator of the club and the club's instructor, ultimately resulted in the instructor's leaving and starting a new club, taking about a half of the original club's members with him. The node colors in Fig. 3 indicate how exactly the two factions ended up splitting.

Fig. 5 presents the results of the entropy centrality-based community detection algorithm, executed with the karate club network. The algorithm discovers the two main club communities, offering a strict refinement of the community structure reported in (Girvan and Newman, 2002) and agreeing with the findings of Medus et al. (2005). For finding the two-community division, 25 iterations of the algorithm were executed with locality $t = 5$. This partition corresponds almost exactly to the actual factions in the club, with an only exception of some "outliers", the nodes with the lowest degree values. The outliers, nodes 5, 10, 11, 12 and 29, were

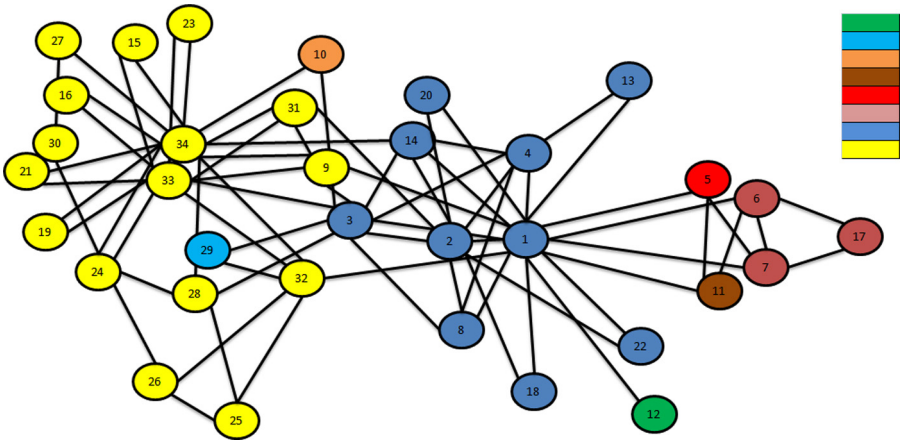

Fig. 5. Entropy centrality algorithm results for karate club problem depicting the sequence of community formation.

Table 2
Clustering algorithms comparison – karate club data (34 nodes, 78 edges).

Number of iterations	Girvan–Newman algorithm			Entropy-based algorithm		
	Clusters	Outliers	Time (s)	Clusters	Outliers	Time (s)
10	1	0	5.6	2	2	0.5
15	2	1	7.4	3	4	0.6
25	4	2	9.5	4	5	0.77

Table 3
Clustering algorithms comparison – football network data (115 nodes, 613 edges).

Number of iterations	Girvan–Newman algorithm			Entropy-based algorithm		
	Clusters	Outliers	Time (s)	Clusters	Outliers	Time (s)
50	1	0	577.5	1	0	205.3
100	3	0	904.4	3	0	388.2
150	6	0	1046.1	7	0	553.1
200	12	3	1094.0	12	3	673.6
250	14	4	1134.7	12	13	738.9

detected first, which is a desirable property for a community detection algorithm that looks to find closely connected groups. Note also that increasing the number of algorithm iterations produces a more granular clusters (perhaps, smaller friendship groups or families), however, any further refinements could not have been validated due to the lack of data.

Table 2 reports the number of clusters identified by Girvan–Newman algorithm (Girvan and Newman, 2002) and the

FloridaState	WashingtonState	FresnoState	Auburn	Tulane	TexasTech
NorthCarolinaState	OregonState	Rice	Alabama	Army	Baylor
Virginia	California	SouthernMethodist	Florida	Cincinnati	Colorado
GeorgiaTech	BrighamYoung	Nevada	Kentucky	AlabamaBirmingham	Kansas
Duke	NewMexico	SanJoseState	Vanderbilt	Akron	IowaState
NorthCarolina	SanDiegoState	TexasElPaso	MississippiState	BowlingGreenState	Nebraska
Clemson	Wyoming	Tulsa	SouthCarolina	Buffalo	TexasA&M
WakeForest	Utah	TexasChristian	Tennessee	Kent	Texas
Maryland	ColoradoState	Hawaii	Mississippi	MiamiOhio	Missouri
SouthernCalifornia	AirForce	VirginiaTech	Georgia	Ohio	OklahomaState
ArizonaState	NevadaLasVegas	BostonCollege	LouisianaState	NorthernIllinois	KansasState
UCLA	NorthTexas	WestVirginia	Arkansas	BallState	Oklahoma
Arizona	ArkansasState	Syracuse	EastCarolina	CentralMichigan	
Washington	BoiseState	Pittsburgh	Houston	EasternMichigan	
Oregon	Idaho	Temple	Louisville	Toledo	
Stanford	NewMexicoState	Rutgers	Memphis		
	UtahState	MiamiFlorida	SouthernMississippi		

Team	Conference
NotreDame	NotreDame
LouisianaTech	LouisianaTech
LouisianaMonroe	LouisianaMonroe
Middle Tennessee State	Middle Tennessee State
LouisianaLafayette	LouisianaLafayette
WesternMichigan	Mid-American
CentralFlorida	CentralFlorida
Marshall	Mid-American
Connecticut	Connecticut
Navy	Navy

Fill color	Conference
	Atlantic Coast
	Pacific Ten
	Mountain West
	Big West
	Big East
	Western Athletic
	Southeastern
	Conference USA
	Mid_American
	Big Twelve

Fig. 6. Clusters for NCAA Division I-A football teams.

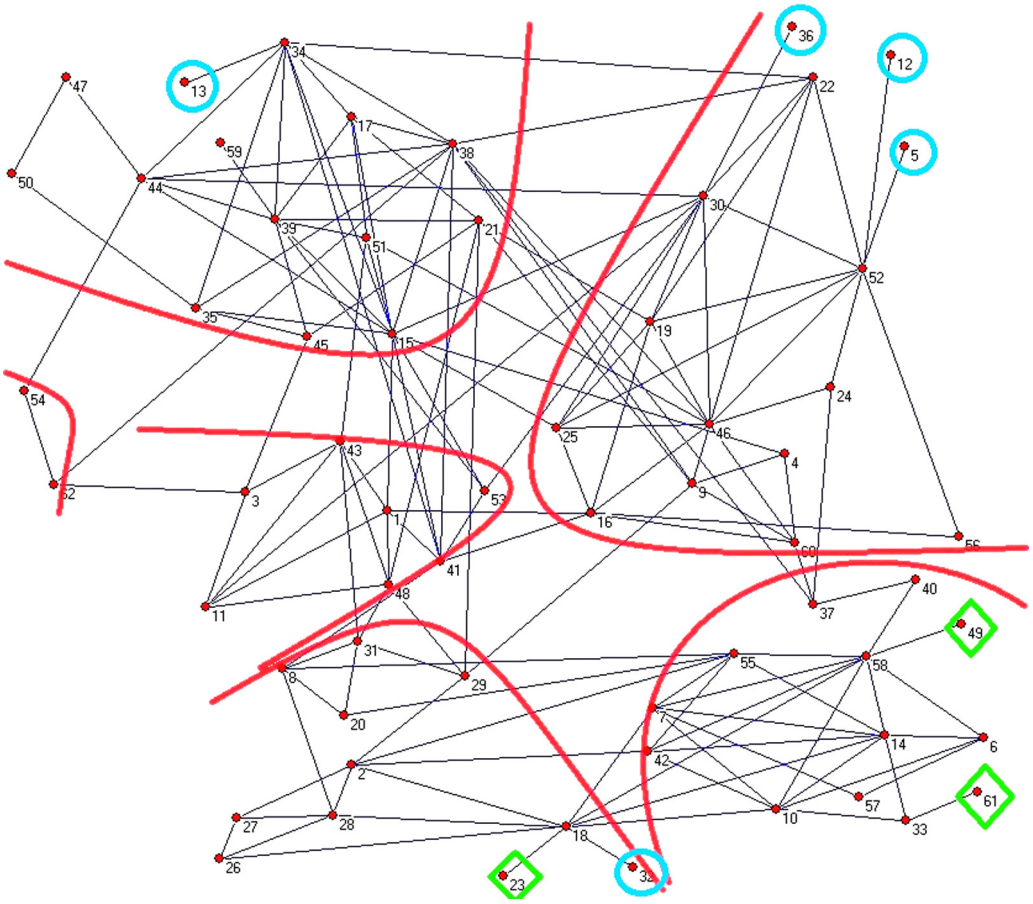

Fig. 7. Clusters for the Dolphin Network.

presented entropy-based algorithm as a function of the number of iterations, together with the respective computational times. In each algorithm, an iteration constitutes a removal of a single edge from the network: in identifying an edge to be removed, the former algorithm computes betweenness centralities for all the edges, while the latter computes entropy centralities for all the nodes and the changes in these centralities that would result from the removal of every edge. Thus, the number of centrality evaluations, required in each iteration of the presented algorithm, is $\mathcal{O}(N)$ times greater than that in the Girvan–Newman algorithm. Yet, the observed algorithm runtimes differ by an order of magnitude in favor of the entropy-based algorithm, due to high efficiency in the computation of entropy centrality. Note that both algorithms are hierarchical, in that they will continue breaking communities apart until the last edge has been removed from the network at hand; an analyst is free to stop this process at any point. In Table 2, and in the tables corresponding to the subsequent experiments, cluster-based metrics are reported for multiple breakpoints in the algorithms' execution, for illustrative purposes. All the experimental results presented in this paper have been obtained using MATLAB version R2012b on a desktop with Intel i3-2120 processor (3.3 GHz, 2 cores) and 8 GB RAM.

3.2. The US Division I football network

Another example is based on the structure of a US college football league (football here is American football, not soccer). The network under study is a representation of the schedule of Division I games in year 2000, with nodes representing teams (identified by their college names) and edges representing regular-season games between the teams they connect. What makes this network interesting is that the true community structure is also available. The teams are divided into conferences containing around 8–12 teams each. Games are more frequent between members of the same conference than between members of different conferences, with teams playing an average of about seven intra-conference games and four inter-conference games in the season. Inter-conference play is not uniformly distributed; teams that are geographically close to one another but belong to different conferences are more likely to play one another than teams separated by large geographic distances (Girvan and Newman, 2002). The entropy centrality based community detection algorithm was applied to this network to identify the conference structure. The algorithm was executed for 200 iterations with transfer locality $t = 5$. The results are presented in Fig. 6. Almost all teams were correctly grouped; a few

Table 4
Clustering algorithms comparison – dolphin network data (62 nodes, 159 edges).

Number of iterations	Girvan–Newman algorithm			Entropy-based algorithm		
	Clusters	Outliers	Time (s)	Clusters	Outliers	Time (s)
45	8	0	39.2	4	16	8.7
75	11	12	48.0	6	22	11.5
100	12	20	51.0	6	31	13.8

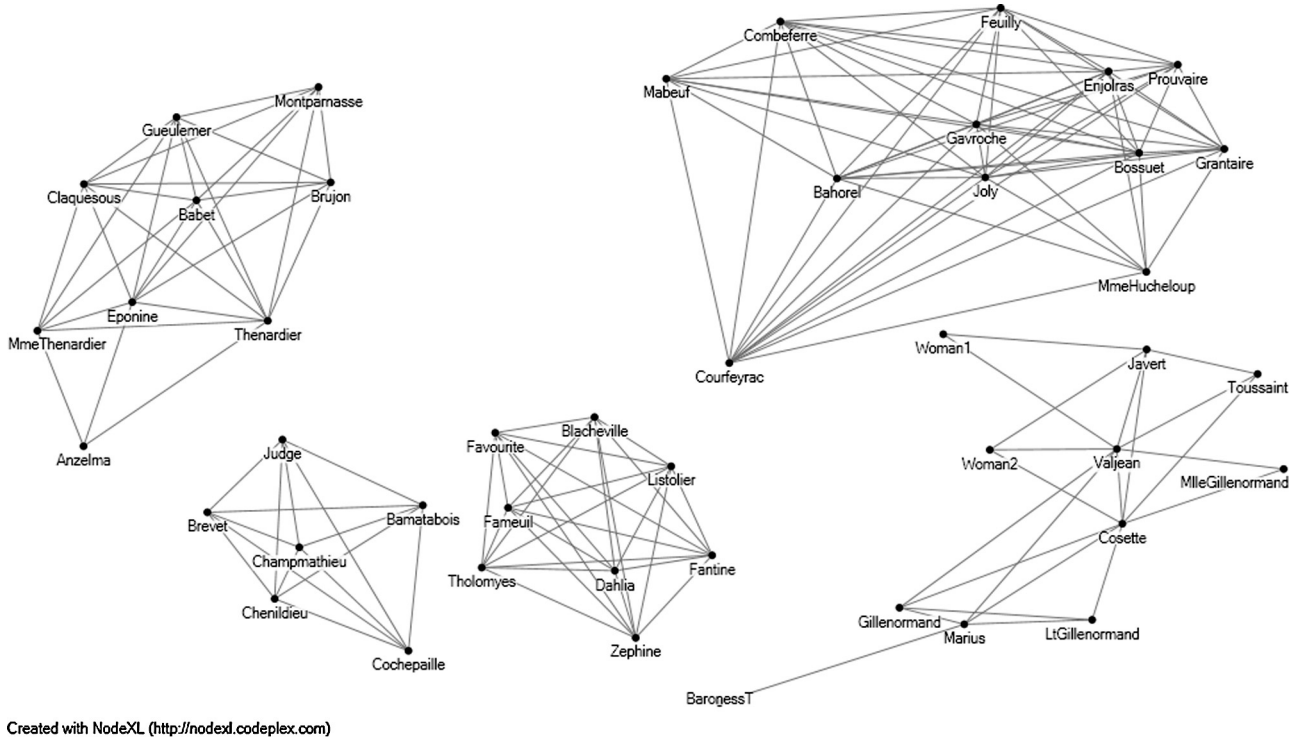

Fig. 8. Clusters for the "Les Miserables" weighted network (100 iterations, 31 outliers).

independent teams that did not belong to any conference were also successfully identified. Overall, only four teams were misclassified: Boise State in actuality belongs in the Western Athletic Conference, Western Michigan and Marshall in the Mid-American Conference, while Utah State is a conference-independent team.

Table 3 offers a comparison of computational times of Girvan–Newman and entropy-based algorithms executed on the football network dataset. Naturally, as the number of iterations increases, the network becomes more sparse. It is worth noting that the number of shortest paths remaining in the network decreases sharply, which explains why an iteration cost in a Girvan–Newman algorithm drops faster than that in the entropy-based algorithm.

3.3. The Dolphin Network

Finally, the algorithm was run on a classic network called "The Dolphin Network" (Lusseau, 2003). The network represents a community of 62 bottlenose dolphins in Doubtful Sound, New Zealand. The algorithm was executed for 45 iterations with locality $t = 5$. The results are presented in Fig. 7 and Table 4: the dolphin community is split into 6 main clusters identified by the entropy centrality community detection algorithm. These results match well with those reported by previously existing clustering algorithms. The singled-out outliers, nodes 5, 12, 13, 32 and 36, were detected first, while nodes 23, 49, 61 were not separated into an isolated cluster. In its current design, the entropy centrality algorithm cannot not be used to distinguish overlapping communities, which is another community detection task typically explored based on this dataset.

Note that the interpretation of entropy centrality emphasizes diversity in multi-way information exchanges between nodes, as opposed to emphasizing connectivity. Therefore, the presented clustering algorithm first and foremost seeks to achieve high cohesion within each discovered group, and this is why outliers tend to be removed early in all the attempted experiments.

4. Discussion and conclusion

This paper introduces a measure of node centrality as the entropy of flow destination in a walk-based transfer process with Markovian property. Entropy centrality can be particularly useful in large social network analysis, where the multitude of paths between node pairs makes the differences in the typically-used betweenness centrality values almost negligible. Entropy centrality is well-interpretable and easy to compute exactly using matrix multiplication.

By design, entropy centrality can be interpreted as a measure of node potential for information spread: the more diverse set of destinations a node can engage, the higher centrality it boasts. Moreover, by adjusting the settings of the Markovian transfer process, one is able to measure entropy centrality at different localities, establishing the value of every node's position with respect to its local neighbors, or globally, with respect to the whole network. The notion of locality in entropy centrality definition is akin to that of reach, used to define reach centrality, however, in the stochastic transfer process context, these two are not quite the same.

Entropy centrality is conducive to quantifying the properties of a serial duplication network flow process, thus taking a particular spot in Borgatti's typology of social network processes/metrics. This observation motivates further investigations into entropy centrality utility for viral marketing studies, where spread of ideas or products takes place simultaneously over multiple network paths or walks. A profit-sharing product distribution strategy where distributors are constantly recruited by the existing distributors directly from the consumer population is a good example of such a potential analysis application.

Entropy centrality can be useful for network visualization, with globally central nodes placed into a canvas first, uniformly spaced, and with the surrounding other nodes becoming more distant as their local centrality drops. Such a visualization would emphasize information exchange capabilities of nodes at multiple levels, instead of relying exclusively on local network structure captured

Table 5
Entropy-based clustering algorithm – Les Miserables data (77 nodes, 254 edges, weighted).

Number of iterations	Entropy-based algorithm		
	Clusters	Outliers	Time (s)
25	2	20	14.6
50	4	25	27.2
100	5	31	47.9
150	4	48	58.8

by its edges. This paper also explores how entropy centrality can be utilized by an iterative algorithm to effectively detect communities. Computational experiments on networks with known cluster ground truth showcase the effectiveness of the presented method.

Additional insights, drawn from the experiments with the entropy centrality-based clustering algorithm, are notable. First, entropy centrality appears to remain informative with weighted networks (the matrix-based way of computing entropy centrality values does not require any significant revision). Fig. 8 and Table 5 give the results and computational times for the presented clustering algorithm application to a weighted dataset of character co-appearances in the text of "Les Miserables": the discovered communities comply with the results previously reported in the literature. Second, it is observed that in its current form, entropy centrality may not be as useful for analyzing directed networks. When applied to prisoner relationship data (67 nodes, 182 edges), the entropy centrality-based clustering algorithm failed to discover well-interpretable clusters. This is due to the fact that in a directed network, many actors have very limited options for information spread, and are isolated as "outliers" early in the algorithm's execution. Also, experiments with larger networks revealed that runtime-wise, the clustering algorithm's applicability has similar limitations as Girvan–Newman algorithm does: more specifically, the former can work with datasets with up to 500 nodes, whereas the latter becomes slow clustering just 100 nodes. Meanwhile, computational efficiency of one-time evaluation of entropy centrality values over all network nodes remains very high, as expected.

On the final note, future research on the use of entropy centrality for social network analysis can focus on evaluating strategic network positions of *groups* of nodes. Having directly computed all the absorption probabilities (i.e., from state $i \in N$ into state $j \in N$) for the Markovian transfer process (i.e., from state $i \in N$ into state $j \in N$), one can search for subsets of strategically positioned nodes at various localities. Other research directions include increasing the computational efficiency of the presented methods, and devising methods for detecting overlapping network communities.

Acknowledgment

This research has been supported in part by the National Science Foundation (Award #62288) and by a Multidisciplinary University Research Initiative (MURI) grant (#W911NF-09-1-0392) for "Unified Research on Network-based Hard/Soft Information Fusion", issued by the US Army Research Office (ARO) under the program management of Dr. John Lavery.

References

Bonacich, P., 1972. Factoring and weighting approaches to status scores and clique identification. J. Math. Sociol. 2 (1), 113–120.
Bonacich, P., Lloyd, P., 2001. Eigenvector-like measures of centrality for asymmetric relations. Soc. Netw. 23 (3), 191–201.
Borgatti, S.P., 2005. Centrality and network flow. Soc. Netw. 27 (5), 5–71.
Estrada, E., Rodriguez-Velazquez, J.A., 2005. Subgraph centrality in complex networks. Phys. Rev. E 71 (5), 056103.
Freeman, L.C., 1979. Centrality in social networks: conceptual clarification. Soc. Netw. 1 (21), 5–239.
Girvan, M., Newman, M.E.J., 2002. Community structure in social and biological networks. Proc. Natl. Acad. Sci. U. S. A. 99 (782), 1–7826.
Lusseau, D., 2003. The emergent properties of a dolphin social network. Proc. Biol. Sci. 270, S186–S188.
Medus, A., Acuna, G., Dorso, C., 2005. Detection of community structures in networks via global optimization. Phys. A 358, 593–604.
Newman, M.E.J., 2005. A measure of betweenness centrality based on random walks. Soc. Netw. 27 (3), 9–54.
Noh, J.D., Rieger, H., 2004. Random walks on complex networks. Phys. Rev. Lett. 92, 118701-1-4.
Stephenson, K., Zelen, M., 1989. Rethinking centrality: methods and examples. Soc. Netw. 11, 1–37.
Tutzauer, F., 2007. Entropy as a measure of centrality in networks characterized by path-transfer flow. Soc. Netw. 29 (2), 249–265.