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.

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/

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.