Knowledge-Based Systems 30 (2012) 136–150

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

A novel measure of edge centrality in social networks

Pasquale De Meo [a], Emilio Ferrara [b,*], Giacomo Fiumara [a], Angela Ricciardello [b]

[a] Department of Physics, Informatics Section, University of Messina, V.le F. Stagno D'Alcontres 31, 98166 ME, Italy
[b] Department of Mathematics, University of Messina, V.le F. Stagno D'Alcontres 31, 98166 ME, Italy

ARTICLE INFO

Article history:
Received 23 October 2011
Received in revised form 6 January 2012
Accepted 6 January 2012
Available online 15 January 2012

Keywords:
Complex networks
Social networks
Centrality measure
Network analysis
Network science

ABSTRACT

The problem of assigning centrality values to nodes and edges in graphs has been widely investigated during last years. Recently, a novel measure of node centrality has been proposed, called κ-path centrality index, which is based on the propagation of messages inside a network along paths consisting of at most κ edges. On the other hand, the importance of computing the centrality of edges has been put into evidence since 1970s by Anthonisse and, subsequently by Girvan and Newman. In this work we propose the generalization of the concept of κ-path centrality by defining the κ-path edge centrality, a measure of centrality introduced to compute the importance of edges. We provide an efficient algorithm, running in $O(\kappa m)$, being m the number of edges in the graph. Thus, our technique is feasible for large scale network analysis. Finally, the performance of our algorithm is analyzed, discussing the results obtained against large online social network datasets.

1. Introduction

In the context of the social knowledge management, Social Network Analysis (SNA) is attracting an increasing attention by the scientific community, in particular during the latest years. One of the main motivations is the unprecedented success of phenomena such as online social networks and online communities. In this panorama, not only from a scientific perspective but also for commercial or strategic motivations, the identification of the principal actors inside a network is very important.

Such an identification requires to define an *importance measure* (also referred to as *centrality*) to weight nodes and/or edges.

The simplest approaches to computing centrality consider only the *local topological properties* of a node/edge in the social network graph: for instance, the most intuitive node centrality measure is represented by the degree of a node, i.e., the number of social contacts of a user. Unfortunately, local measures of centrality, whose esteem is computationally feasible even on large networks, do not produce very faithful results [1].

Due to these reasons, many authors suggested to consider the *whole social network topology* to compute centrality values. A new family of centrality measures was born, called *global measures*. Some examples of global centrality measures are *closeness* [2] and *betweenness centrality* (for nodes [3], and edges [4,5]).

Betweenness centrality is one of the most popular measures and its computation is the core component of a range of algorithms and applications. Betweenness centrality relies on the idea that, in social networks, information flows along *shortest paths*: as a consequence, a node/edge has a high betweenness centrality if a large number of shortest paths crosses it.

Some authors, however, raised some concerns on the effectiveness of betweenness centrality. First of all, the problem of computing the exact value of betweenness centrality for each node/edge of a given graph is computationally demanding – or even unfeasible – as the size of the analyzed network grows. Therefore, the need of finding fast, even if approximate, techniques to compute betweenness centrality arises and it is currently a relevant research topic in Social Network Analysis.

A further issue is that the assumption that information in social networks propagates only along shortest paths could not be true [6]. By contrast, information propagation models have been provided in which information, encoded as messages generated in a source node and directed toward a target node in the network, may flow along *arbitrary* paths. In the spirit of such a model, some authors Newman [7], Noh and Rieger [8] suggested to perform random walks on the social network to compute centrality values.

A prominent approach following this research line is the work proposed in [9]. In that work, the authors introduced a novel node centrality measure known as *κ-path centrality*. In detail, the authors suggested to use self-avoiding random walks [10] of length κ (being κ a suitable integer) to compute centrality values. They provided an approximate algorithm, running in $O(\kappa^3 n^{2-2\alpha} \log n)$ being n the number of nodes and $\alpha \in \left[-\frac{1}{2}, \frac{1}{2}\right]$.

In this paper we extend that work [9] by introducing a measure of *edge centrality*. This measure is called *κ-path edge centrality*. In our approach, the procedure of computing edge centrality is

* Corresponding author.
E-mail addresses: pdemeo@unime.it (P. De Meo), eferrara@unime.it (E. Ferrara), gfiumara@unime.it (G. Fiumara), aricciardello@unime.it (A. Ricciardello).

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2012.01.007

viewed as an *information propagation problem*. In detail, if we assume that multiple messages are generated and propagated within a social network, an edge is considered as "central" if it is frequently exploited to diffuse information.

Relying on this idea, we simulate message propagations through random walks on the social network graphs. In our simulation, in addition, we assume that random walks are *simple* and of *bounded length* up to a constant and user-defined value κ. The former assumption is because a random walk should be forced to pass no more than once through an edge; the latter, because, as in [11], we assume that the more distant two nodes are, the less they influence each other.

The computation of edge centrality has many practical applications in a wide range of contexts and, in particular, in the area of knowledge-based (KB) systems. For instance in KB systems in which data can be conveniently managed through graphs, the procedure of weighting edges plays a key role in identifying *communities*, i.e., groups of nodes densely connected to each other and weakly coupled with nodes residing outside the community itself [12,13]. This is useful to better organize available knowledge: think, for instance, to an e-commerce platform and observe that we could partition customer communities into smaller groups and we could selectively forward messages (like commercial advertisements) only to groups whose members are actually interested to them. In addition, in the context of Semantic Web, edge centralities are useful to quantify the strength of the relationships linking two objects and, therefore, it can be useful to discover new knowledge [14]. Finally, in the context of social networks, edge centralities are helpful to model the intensity of the social tie between two individuals [15]: in such a case, we could extract patterns of interactions among users in virtual communities and analyze them to understand how a user is able to influence another one. The main contributions of this paper are the following:

- We propose an approach based on *random walks* consisting of up-to κ edges to compute *edge centrality*. In detail, we observe that many approaches in the literature have been proposed to compute node centrality but, comparatively, there are few studies on edge centrality computation (among them we cite the edge betweenness centrality introduced in the Girvan–Newman algorithm [5]). In addition, Newman [7], Noh and Rieger [8], Brandes and Fleischer [16] successfully applied random walks to compute node centrality in networks. We suggest to extend these ideas in the direction of edge centrality, and, therefore, this work is the *first attempt* to compute edge centrality by means of random walks.
- We design an algorithm to efficiently compute edge centrality. The worst case time complexity of our algorithm is $O(\kappa m)$, being m the number of edges in the social network graph and κ a constant (and typically small) factor. Therefore, the running time of our algorithm scales in *linear fashion* against the number of edges of a social network. This is an interesting improvement of the state-of-the-art: in fact, *exact algorithms* for computing centrality run in $O(n^3)$ and, with some ingenious optimizations they can run in $O(nm)$ [17,5]. Unfortunately, real-life social networks consist of up to millions nodes/edges [18], and, therefore these approaches may not scale well. By contrast, our algorithm works fairly well also on large real-life social networks even in presence of limited computing resources.
- We provide results of the performed experimentation, showing that our approach is able to generate reproducible results even if it relies on random walks. Several experiments have been carried out in order to emphasize that the κ-path edge centrality computation is feasible even on large social networks. Finally, the properties shown by this measure are discussed, in order to characterize each of the studied networks.

The paper is organized as follows: in Section 2 we provide some background information on the problems related to centrality measures. Section 3 presents the goal of this paper and our κ-path edge centrality, including the fast algorithm for its computation. The experimental evaluation of performance of this strategy is discussed in Section 4 and some possible applications of our approach are presented in Section 5. Thus, the paper concludes in Section 6.

2. Background about centrality measures and applications

In this section we review the concept of centrality measure and illustrate some recent approaches to compute it.

2.1. Centrality measure in social networks

One of the first (and the most popular) node centrality measures is the *betweenness centrality* [3]. It is defined as follows:

Definition 1 (*Betweenness centrality*). Given a graph $G = \langle V,E\rangle$, the betweenness centrality for the node $v \in V$ is defined as

$$C_{B_n}(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1}$$

where s and t are nodes in V, σ_{st} is the number of shortest paths connecting s to t, and $\sigma_{st}(v)$ is the number of shortest paths connecting s to t passing through the node v.

If there is no path joining s and t we conventionally set $\frac{\sigma_{st}(v)}{\sigma_{st}} = 0$.

The concept of centrality has been defined also for the edges in a graph and, from a historical standpoint, the first approach to compute edge centrality has been proposed in 1971 by Anthonisse [4,19] and was implemented in the GRADAP software package. In this approach, edge centrality is interpreted as a "flow centrality" measure. To define it, let us consider a graph $G = \langle V,E\rangle$ and let $s \in V$, $t \in V$ be a *fixed* pair of nodes. Assume that a "unit of flow" is injected in the network by picking s as the source node and assume that this unit flows in G along the shortest paths. The *rush index* associated with the pair $\langle s,t\rangle$ and the edge $e \in E$ is defined as

$$\delta_{st}(e) = \frac{\sigma_{st}(e)}{\sigma_{st}}$$

being, as before, σ_{st} the number of shortest paths connecting s to t, and $\sigma_{st}(e)$ the number of shortest paths connecting s to t passing through the edge e. As in the previous case, we conventionally set $\delta_{st}(e) = 0$ if there is no path joining s and t.

The rush index of an edge e ranges from 0 (if e *does not belong* to any shortest path joining s and t) to 1 (if e belongs to *all* the shortest paths joining s and t). Therefore, the higher δ_{st}, the more relevant the contribution of e in the transfer of a unit of flow from s to t. The *centrality* of e can be defined by considering all the pairs $\langle s,t\rangle$ of nodes and by computing, for each pair, the rush index $\delta_{st}(e)$; the centrality $C_{R_e}(e)$ of e is the sum of all these contributions

$$C_{R_e}(e) = \sum_{s \in V} \sum_{v \in V} \delta_{st}(e)$$

More recently, in 2002, Girvan and Newman proposed in [5] a definition of *edge betweenness centrality* which strongly resembles that provided by Anthonisse.

According to the notation introduced above, the edge betweenness centrality for the edge $e \in E$ is defined as

$$C_{B_e}(e) = \sum_{s \neq t \in V} \frac{\sigma_{st}(e)}{\sigma_{st}} \tag{2}$$

and it differs from that of Anthonisse because the source node s and the target node t must be different.