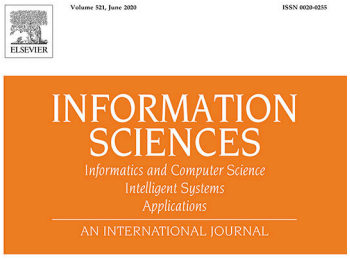
Journal Pre-proofs



A novel measure to identify influential nodes: Return Random Walk Gravity Centrality

Manuel Curado, Leandro Tortosa, Jose F. Vicent

PII: S0020-0255(23)00093-2
DOI: https://doi.org/10.1016/j.ins.2023.01.097
Reference: INS 18628

To appear in: *Information Sciences*

Received Date: 31 October 2022
Accepted Date: 15 January 2023

Please cite this article as: M. Curado, L. Tortosa, J.F. Vicent, A novel measure to identify influential nodes: Return Random Walk Gravity Centrality, *Information Sciences* (2023), doi: https://doi.org/10.1016/j.ins.2023.01.097

A novel measure to identify influential nodes: Return Random Walk Gravity Centrality

Manuel Curado[a,*], Leandro Tortosa[b], Jose F. Vicent[b]

[a]*Polytechnic School, Catholic University of Murcia, Campus Los Jeronimos, s/n, E-30107 Murcia, Spain*
[b]*Department of Computer Science and Artificial Intelligence, University of Alicante, Campus de San Vicente, Ap. Correos 99, E-03080, Alicante, Spain*

Abstract

To identify influential nodes in real networks, it is essential to note the importance of considering the local and global information in a network. In addition, it is also key to consider the dynamic information. Accordingly, the main aim of this paper is to present a new centrality measure based on return random walk and the effective distance gravity model (C_{RRWG}). This new metric increases the relevance of nodes with a dual role: i) at the local level, they are important in their community or cluster, and ii) at the global level, they give cohesion to the network. It has advantages over other traditional models of centrality since it considers the global and local information, as well as the information of the dynamic interaction between the nodes, as recent studies on community-aware centrality measures demonstrate. Thus, the combination of dynamic and static information makes it easier to detect influential nodes in complex networks. To validate the effectiveness of the proposed centrality measure, it is compared with classic measures, such as Degree, Closeness, Betweenness, PageRank, and other measures based on the gravity model, effective distance and community-aware approaches. The experimental results show the effectiveness of C_{RRWG} through a set of experiments on different types of networks.

Keywords: centrality measure, effective distance, random paths, densification, gravity model

1. Introduction

The analysis of complex networks has been the focus of much research interest in recent years [1]. A large number of real problems can be modelled using network theory, such as transport networks, urban networks or social networks[2, 3]. It can be said that the Complex Network Theory allows us to understand, and attempt to solve, a wide range of real systems, from technological networks to social networks, including disease control, the spreading of rumours, biological systems, social systems, time series predictions or the propagation of information, among others [4, 5, 6].

With the advent of new information technologies, the ability to generate data has increased considerably, representing a phenomenon that, if studied and analysed appropiately, can contribute provide important advances to science. When the data can be divided into related parts, they can be represented by a graph structure. This structure has been widely used to manage data that resides in irregular domains and whose applications are highly diverse and involve topics as diverse as the treatment of energy transmission networks, vehicle and people flow networks, social networks, unsupervised classification or high-dimensional

*Corresponding author
Email addresses: mcurado@ucam.edu (Manuel Curado), tortosa@ua.es (Leandro Tortosa), jvicent@ua.es (Jose F. Vicent)

data analysis, among others [7, 8]. Due to its enormous impacts on our daily lives, the identification of influential nodes in all types of complex networks (see [9]) has become a significant research area that plays an important role [10] in aspects related to both structure and functionality [11]. For instance, in social networks, the central nodes indicate influential people [12]; in urban networks, these nodes indicate locations where the flow of pedestrians or traffic is important [13]; and in information or the spread of disease, they indicate influential disseminators of information or diseases [14].

Despite the complexity of the definition of the concept of node centrality, many methods of evaluating the influence of a node in a network have been proposed [15]. Some of these metrics do not take into account global information, focusing, rather, on local information. Among these are Degree Centrality (DC) [16], Pagerank (PG) [17], the original K-Shell [18] or the eigenvector centrality (EC) [15]. However, other proposed centralities fundamentally depend on the connectivity in the network, such as betweenness [19] or closeness (CC) [20]. Although these popular metrics perform relatively well, they do have limitations. In this way, DC considers the importance of the neighbour (ignoring the global influence). The PG works successfully on directed networks but not on non-directed ones. CC and BT are highly complex measures that are responsive to the topology of the network. However, above all, these classic methods consider either local or global information, but not both.

In an attempt to alleviate the division between measures of centrality focused on local information or others that cover global information [19, 21], new measures have been proposed. A review of the state of the art, clarifying concepts and metrics, classifying problems and methods, is carried out in [22]. In [23], the authors perform different rankings of the nodes of a network, taking into account the topology and the data associated with the nodes. In [24] and [25], two round-trip centralities, which reinforce dense networks, are presented. Furthermore, a model based on the iterative allocation of resources is proposed in [26]. Deng et al. [27] use the inverse square law to detect vital nodes. Li et al. [11] propose a measure based on Newton's law of gravity model. In addition, other methods have been proposed to identify central nodes, such as the method based on random walks [19], a model based on entropy [28] or a centrality based on quasi-Laplacian energy of networks [29].

Based on the Newton's law, a number of metrics have been proposed to select important nodes in complex networks. These proposed metrics consider that influence of the nodes depends not only on the direct neighbours but also on the nodes that are in a distance radius, so that the influence of the node can be seen as the sum of the attraction with the other nodes. In [27], the authors proposed a model that considers the degree of the node as the mass and the shortest topological path as its distance. The calculation of the shortest paths is computationally very expensive, and thus the truncation radius was established [30]. The problem is that a significant number of complex real networks have a dynamic topological structure that is not clearly visible and which contains information that better identifies vital nodes. In this context, the effective distance [31] helps to solve the problem of discovering hidden dynamic information. Thus, Shang et al. [32] presented an effective distance based on Newton's Law to measure the separation of two nodes.

Real networks often exhibit structures of groups of nodes, called communities, whose definition differs depending on context and assumptions, resulting in a wide variety of characterizations. Thus, for instance, in [33], the authors focus on models that generate communities in complex networks as well as on the development of a basis for community detection algorithms (using the maximization of modularity as a basis for community detection).

Meanwhile, classic centrality measures are often unable to distinguish between nodes integrated into similar network regions, making it necessary to incorporate community structure information. To address this problem, researchers have developed community-aware centralities that classify nodes in terms of intra-community and inter-community ties, taking into account local and global information [34, 35]. Thus, for example, community-based betweenness centrality considers only the shortest paths with endpoints in different communities [36]. The identification of influential spreaders in networks, considering both the number and the size of the communities that are directly linked by a node, is developed in the Community-based Centrality (CBC) [37], although this has the disadvantage of its similarity to normalized DC when the number of communities is high. Another example is the Community-based Mediator (CBM), which ranks nodes depending on the number of connections they have with other communities, weighted by the relative sizes of the communities [38]. In this metric, if the external and internal link density of a node are equal, it

reduces the degree centrality. The K-Shell with community centrality (CKS), developed by [39], determines influential nodes embedded in local and global communities but has the problem of the redundancy of nodes existing at the same hierarchical level. Comm Centrality (COMM) is presented in [40], being a centrality measure that requires information only at the community level to identify important nodes. It searches for hubs and bridges while considering the strength of the structure. However, this metric has the problem of not considering isolated communities.

The Community hub-bridge (CHB) metric calculates the importance of a node as the sum of its intra-community and inter-community links, weighted by the size of the community of the node and the number of communities to which it connects [41]. One of the drawbacks of this measure is that it has a large number of clusters. The nodes that act as bridges between communities, belonging to several of them, have a great influence despite their low degree, occurring mostly in networks with overlapping community structures. Very recently, the vitality of modularity was proposed in [42] (Modularity Vitality Centrality, CMV). It generalizes overlapping communities based on the community detection approach, known as modularity, but has limitations, such as the resolution limit and the local maximum.

An extensive comparative analysis of several community-aware centrality measures was performed in [35], reporting on the best strategy to use according to the community structure presented by the network under study. Moreover, there exist walk-based centrality measures, such as [43], which focus on the local representation of a graph by using user-based communities with centrality-aware random walks. The main problem of random walk-based centrality studies, such as [44], where the centrality of a node is closely related to the notion of navigability, or [45], where the authors proposed a metric that eliminates backtracking walks and that can be interpreted as standard Katz on a modified network, are that they are only focused on either local or global information. A general algorithm has not been defined for all kinds of networks. In a first attempt to solve this problem, in [24, 25], the authors define a general metrics based on Return Random Walk link predictor [46] with the aim being to obtain local and global information in dense networks.

Thus, the main contribution of this study is the proposal of a new centrality measure that focuses on increasing the relevance of the relevance of nodes with a dual role: at the local level, they are important in their community or cluster, and at the global level, they give cohesion to the network. It combines the following approaches: i) a walk-based link predictor named Return Random Walk, ii) effective distances, and iii) a gravity model. The strength of this approach is the detection of important nodes in complex networks, combining static and dynamic information from the use of effective distances in a gravity model. The use of return random walks lets us highlight the community structure, considering both global and local information, increasing the centrality of the relevant nodes in the task of reinforcing its community.

To check the robustness of the model presented, we have used seven real networks (directed, undirected, weighted and unweighted). By means of the Susceptible-Infected (SI) model, we test the speed and the reachability of the diffusion, comparing our proposal with seven centrality measures. The results obtained evidence the improvement of the proposed method.

To achieve the aims proposed in the paper, the rest of the article is structured as follows. With the aim of evaluating the proposal, various well-known centrality measures are presented in Section 2. Section 3 describes the methodology used in developing the metric that helps us to identify influential nodes. To validate the model, it is necessary to obtain results and compare them with other models; this is done in Section 4. Finally, the conclusions are presented in Section 5.

2. Preliminaries

The proposed measure of centrality is based on random walk betweenness centrality and the gravitational model with effective distance. To validate the effectiveness of the measure, using the Susceptible-Infected (SI) model, we perform a comparison with various measures of centrality. This comparison is conducted for directed and undirected networks and networks with and without weights.
To understand the methodology of the new model, some preliminary comments are required.

2.1. Centrality measures

Given a graph $G(V, E, W)$, where V is the set of N nodes, E is the set of edges, W is a weighted matrix and A is the adjacency matrix, the processes of calculating different centrality measures and models are described in Table 1 for classic centrality measures and in Table 2 for community-aware centrality measures.

Centrality	**Definition**	**Formula**
Degree	Neighbours where a node is connected to [16]. Where V is the set of nodes and a_{ij} is the element ij of the adjacency matrix A.	$K_i = \sum_{j \in V} a_{ij}.$
Closeness	This is the inverse of the sum of the length of the shortest paths [20]. Where d_{ij} is the shortest path distance between node i and node j.	$C_B(i) = \frac{1}{\sum_{j \neq i} d_{ij}}.$
Betweenness (Random-walk)	Number of times that a random walk passes through a node [19]. Where σ_{ij} is the number of random walks from node i to node j, and σ_{itj} represents the number of random walks from node i to node j through node t.	$B_{RW}(i) = \sum_{i \neq t} \sum_{j \neq i,t} \frac{\sigma_{itj}}{\sigma_{ij}}.$
Pagerank	This assigns a numerical weighting to each node, [17]. Where N_i represents the neighbours of node i, α is the damping factor, k_j is the number of nodes to which the node j points, and t represents an iterative parameter.	$PR_i^t = \frac{1-\alpha}{n} + \alpha \sum_{j \in N_i} \frac{PR_j^{t-1}}{k_j}.$

Table 1: Definition and formula of the classic centrality measures.

Centrality	Definition	Formula
Community-based Centrality (CBC)	This weights the node's intra-cluster and inter-cluster links by the size of their clusters [37]. Where C is the total number of clusters, n_{c_q} is the total number of nodes in cluster c_q and k_{i,c_q} is the total number of links of a node i in cluster c_q.	$CBC(i) = \sum_{q=1}^{C} k_{i,c_q}(\frac{n_{c_q}}{N})$.
Community Hub-Bridge (CHB)	This measure weights the node's intra-cluster and inter-cluster links targeting hubs and bridges simultaneously [41]. Where n_{c_q} is node i's cluster size and $\bigvee_{j\in c_l} a_{ij} = 1$ if is connected to, at least one node j in cluster c_l.	$CHB(i) = n_{c_q} \times k_i^{intra} + \sum_{c_l \in C}^{N} \bigvee_{j\in c_l} a_{ij} \times k_i^{inter}$.
Community-based Mediator (CBM)	This is based on the entropy of the inter-cluster and inter-cluster links of a node [38]. Where H_i is the entropy of node i based in the node's ratio of intra and inter-cluster.	$CBM(i) = H_i \times \frac{k_i^{tot}}{\sum_{i=1}^{N} k_i^{tot}}$.
Comm Centrality (COM)	This preferentially targets bridges [40]. Where μ_{c_q} is the fraction of inter-cluster links over the total cluster links in cluster c_q, $\chi = \frac{k_i^{intra}}{max_{(j\in c)} k_j^{intra}} \times R$ and $\varphi = \frac{k_i^{inter}}{max_{(j\in c)} k_j^{inter}} \times R$.	$COMM(i) = (1+\mu_{c_q}) \times \chi + (1-\mu_{c_q}) \times \varphi^2$.
K-Shell with Community (CKS)	This splits network G into two networks. One is made of the nodes and their intra-cluster links, and the other has the nodes and inter-cluster links [39]. Where α^{intra} and α^{inter} are the K-shell value of node i.	$CKS(i) = \delta \times \alpha^{intra}(i) + (1-\delta) \times \alpha^{inter}(i)$.
Modularity Vitality (CMV)	This differentiates a hub from a bridge based on Newman's modularity [42]. Where $Q(G)$ is the network modularity and $Q(G \setminus \{i\})$ is the network's modularity after removal of node i.	$CMV(i) = Q(G) - Q(G \setminus \{i\})$.

Table 2: Definition and formula of the community-aware centrality measures

.

2.2. Effective distances gravity model (EffG)

Inspired by the law of gravitation in physics and based on the idea of attraction between different nodes in a network, Li et al. [47] proposed a new metric to measure the influence of each node. In this method, a node with many neighbours that are close to many nodes is more influential. Thus, the influence of a node

i can be estimated as:

$$C(i) = \sum_{j=1;j\neq i}^{n} \frac{K_i K_j}{d_{ij}^2},$$

where K_i and K_j are the degrees of the nodes i and j, respectively, and d_{ij} is the shortest topological distance between i and j, with $i \neq j$.

This method has two main drawbacks: on the one hand, the computational time for calculating the shortest distances between all pairs of nodes is high. On the other hand, it is difficult for a node to influence other nodes that are at a great topological distance, since the influence decreases step by step and is disturbed by the accumulated noise. To avoid these weaknesses, a gravity model with the introduction of a truncation radius R was introduced. It is based on the idea that only the interactions by node pairs within the truncation radius are considered.

Then, the gravity model with truncation radius R provides a metric where the centrality of a node i is given by

$$C_R(i) = \sum_{d_{ij}\leq R;j\neq i}^{n} \frac{K_i K_j}{d_{ij}^2}.$$

In this model, in addition to considering the change in mass per degree, it can be considered the effective distance [48]. This measure $D_{i|j}$ from node j to node i is defined as:

$$D_{i|j} = 1 - log_2\left(\frac{a_{ji}}{K_j}\right).$$

The idea of incorporating dynamic information containers in the topology of the network motivates the definition of the effective distance gravity model (EffG). Therefore, the EffG centrality score of a particular node i is given by the expression

$$C_{EffG}(i) = \sum_{j=1,j\neq i}^{N} W_{interaction}(i,j) = \sum_{j=1,j\neq i}^{N} \frac{K_i K_j}{D_{j|i}^2},$$

where N in the number of nodes and $W_{interaction}(i,j)$ represents the specific interaction scores between all pairs (i,j) of nodes.

2.3. Susceptible-Infected Model (SI)

To evaluate the efficiency of the proposed measure, the Susceptible–Infected (SI) model is used. It simulates the epidemic spreading process on real-world networks to evaluate the diffusion power of top M nodes.

The classic SI compartmental model divides a population of nodes into two steps or states: susceptible and infected. An infected node cannot return to the susceptible step. The number of infected nodes determines the speed and reachability in the diffusion of the initial infected nodes (or most significant nodes in this study). The model describes how the number of individuals in each of two classes changes with time. In this model, $S(t)$ is the proportion of the population susceptible at time t and $I(t)$ is the infectious node. The SI system is defined by

$$\begin{aligned} \frac{dS}{dt} &= -bSI, \\ \frac{dI}{dt} &= bSI. \end{aligned} \tag{1}$$

Where b is the rate at which the disease is transmitted when an infected node interacts with the susceptible population. Equation 1 describes how individuals move from the susceptible group to the infected group and the second equation describes how the infected group increases.

2.4. Return Random Walk (RRW)

In [46], a link prediction method is defined that infers new intra-cluster links in images, minimizing the noise of inter-cluster edges, based on Dirichlet densification and Return Random Walks (RRW). This method lets us calculate the probabilities of all transitive random paths between two nodes n and m traversing other different transition nodes t and l (see Fig. 1).

In general, this model is capable of calculating the centrality of directed and undirected networks, as well as weighted or unweighted networks. This means that the degree of a node is calculated differently, depending on the type of network studied. Thus, before describing the proposed model, the definition of the degree of a node i must be explained.

If the network is unweighted, we have an adjacency matrix A associated with the graph as

$$A = (a_{ij}) = \begin{cases} 1 & \text{if } e_{ij} \in E \\ 0 & \text{otherwise.} \end{cases}$$

If the network is weighted, there then exists a weighted matrix W defined as

$$W = (w_{ij}) = \begin{cases} w_{ij} & \text{if } e_{ij} \in E \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, if the network is directed, the nodes have an in-degree and an out-degree.

Then, taking into account the above considerations, for undirected and weighted networks, the degree is

$$K_n = \sum_{m \neq n, m=1}^{N} w_{mn}. \tag{2}$$

If the network is unweighted, then $w_{ij} = 1$, for all nonzero elements in W.

In directed and weighted networks, we calculate in-degrees K^- and out-degrees K^+ for a node i using these expressions:

$$K_i^- = \sum_{m \neq i, m=1}^{N} w_{im}, \qquad K_i^+ = \sum_{m \neq i, m=1}^{N} w_{mi}. \tag{3}$$

From expressions (2) and (3), we can define the two-step random walks $P_{n|m}$, reaching a node n from node m through any transition node t (in red colour), and return from n to m through a different transition node l ($t \neq l$) (in blue colour).

For undirected networks,

$$P_{n|m} = \frac{w_{mt} w_{tn}}{K_m K_n} \frac{w_{nl} w_{lm}}{K_n K_m}, \tag{4}$$

while for directed networks

$$P_{n|m} = \frac{w_{mt} w_{tn}}{K_m^- K_n^+} \frac{w_{nl} w_{lm}}{K_n^- K_m^+}. \tag{5}$$

Finally, of all the 4-length paths from node m to node n, we select the one with the maximum probability.

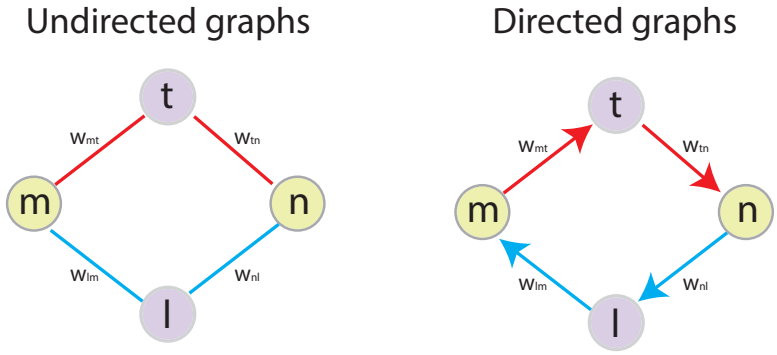


Figure 1: Return Random Walk (RRW) for directed and undirected graphs (for all 4 different nodes m, n, t and l of the network). Green nodes are the origin (m) and destination (n), purple nodes are the transition nodes (t and l), and red links belong to the go path and blue links belong to the return path. In the case of unweighted networks, $W_{ij} = 1$ if exists links between nodes i and j, and 0 otherwise

.

The mathematical explanation of this approach is supported by the knowledge of the powers of the adjacency matrix. The calculation of this matrix squared gives us the final number of paths between its nodes of length two. Following this process, a high power of these matrices lets us obtain the same path length equivalent to the order of the selected power.

In this method, for the adjacency matrix A, we focus on matrix A^4, which indicates the total number of walks (or paths) of length four between all distinct nodes of the graph. For instance, element $A^4(m,n)$ lets us to obtain the number of paths from the node m to n with length four. Applying this argumentation, element $A^4(m,n)$ gives us the number of cycles from m to m of length 4. Therefore, the diagonal of A^4 yields the total number of 4-cycles associated with each node.

Note that cycles of length four are chosen because a balance must be established between the path constraints of the defined random walk and the computational efficiency. The choice of longer cycles does not yield significant variations in terms of node ranking but it does substantially increase the computational time.

In the RRW measure, we seek 4-cycles with a specific constraint: all the vertices belonging to any path should be distinct from other nodes. In Fig. 1 (right), we have that $A^4(m,n) = 1$, so the number of cycles from node m of length 4 is equal to 1. It is worth noting that, in general, we need the paths where all the nodes are different.

It should be pointed out that this calculation works for both directed and undirected networks. In weighted networks, we assign, as the weight of the cycle, the sum of the four weights of the edges. We can then conclude that if $P_{n|m}$ represents the two-step random walk from node m to n and returning n with different transition node, $P_{n|m}$ is the cycle of $A^4(m,n)$, where the four nodes of the cycle are different.

In [24], we adapt this approach to obtain a new metric based on random walk betweenness to analyse different problems using real-world undirected networks, such as the spreading of a virus, the influence of a social account in a social network, the importance of a character in the connectivity of a story, or to detect and minimize the transmission of information in a terrorist network. Moreover, in [25], the authors show that this method is valuable to study the behaviour of the transition nodes in different directed networks such as the transport mobility of people or the importance of different elements in chemical compounds. Through the definition of a centrality model based on four values or indices, it is possible to analyse the density, the strength of the links, or the importance of the nodes in the network based on local and global information. Although these methods capture the importance of nodes in dense clusters, further action is required in order to consider all the dynamic structures of the network.

Considering the components Return Random-Walk, Effective Distance and Gravity Model, and taking into account the community structure information (local, global, static and dynamic) for all kinds of networks (directed, undirected, weighted and unweighted) a new centrality measure is proposed. For this purpose, a formal analysis of the methodology used is described in detail in the next section. Furthermore, to

validate the efficiency of the model, a comparison with other measures of centrality is performed using the Susceptible-Infected (SI) epidemiological model.

3. Methodology

We propose a new centrality measure combining three concepts: i) the effective distances to capture the static and dynamic information of the topology of a network, ii) the gravity model to measure the power of attraction of a node in the network, and iii) return random walks to detect the real strength of each node in a community context.

3.1. Formal analysis

The identification of influential nodes in real networks has been widely studied in recent decades. Recent studies highlight the importance of considering not only the local (intra-cluster) or the global information (inter-cluster) of a network. These works underline the salience of also considering the dynamic information of the network since it provides important information on the effective interaction between nodes. Recently, in [32], the combination of effective distances and the effect of the gravity model was shown to be efficient in capturing the dynamic interactions of a network.

Our proposal, in addition to capturing the dynamic interactions of a network, detects the nodes that strengthen the network (through which a significant amount of information passes), by focusing on the importance of dense clusters.

This proposal works for both directed and undirected networks, with the only difference of applying Equations 4 or 5. The methodology can then be explained in four steps:

3.1.1. Step 1: Calculation of the return random walks between all pairs of nodes.

In this step, we calculate the return random walks between all pairs of nodes. Let us assume that we have c different paths in the return random walks $P^i_{n|m}$, that is

$$P^i_{n|m} = \{P^1_{n|m}, P^2_{n|m}, P^3_{n|m}, \dots P^c_{n|m}\}, \tag{6}$$

where $0 \leq i \leq c$.

3.1.2. Step 2: Calculation of the effective distance between all nodes.

The effective distance gravity model (EffG) is based on the effective distance proposed in [48]. However, the probability used in this model is insufficient to capture all the information of a dynamic network, since it gives excessive importance to the nodes that play the role of inter-cluster node. To address this problem, the use of the return random walks to calculate new effective distances is proposed. The effective distances are

$$D^i_{n|m} = 1 - log_2(P^i_{n|m}), \quad i = 1, \dots, c. \tag{7}$$

Of all the effective distances obtained by Equation 7, the lowest is selected

$$D_{n|m} = min\{D^i_{n|m}\}, \quad i = 1, \dots, c. \tag{8}$$

Note that the effective distances satisfy $D_{n|m} \neq D_{m|n}$ and $D_{n|n} = 0$.

3.1.3. Step 3: Calculation of the interaction between all pairs of nodes.

With the aim of studying the local and global information of the network, the gravity model values between all pairs of nodes considering the effective distance are

$$W_{interaction}(m,n) = \frac{K_m K_n}{(D_{n|m})^2}, \tag{9}$$

where $W_{interaction}(m,n)$ is the interaction score between a pair of nodes n and m, K_n and K_m are the degree of the nodes and $D_{n|m}$ is the effective distance from node n to node m.

It is worth noting that Equation 9 is for undirected networks; in the case of directed networks the equation is

$$W_{interaction}(m,n) = \frac{K_m^- K_n^+}{(D_{n|m})^2}, \tag{10}$$

where $D_{n|m}$ is calculated by the expression Equation 8.

3.1.4. Step 4: Calculation of the centrality of each node.

Finally, the centrality values (RRW_{ED}) of a node i are obtained by summing the interaction scores of all remaining nodes.

$$C_{RRWG}(i) = \sum_{j=1, j\neq i}^{N} W_{interaction}(i,j) = \sum_{j=1, j\neq i}^{N} \frac{K_i K_j}{(D_{j|i})^2}. \tag{11}$$

As a result, a ranking vector named Return Random Walk Gravity Centrality (C_{RRWG}) is obtained, denoting the centrality measure value of the nodes of the network.

3.2. Calculation of the centrality measure: a toy sample

In order to clarify the methodology used to calculate the centrality, a toy example is presented. Thus, in Figure 2 we show a directed weighted network with 6 nodes and 7 edges. We calculate the centrality values following the four steps described in Section 3.1.

4. Experimental Section

In this section, we validate the effectiveness of the proposal (C_{RRWG}) by comparing with centralities such as Degree, Closeness, Betweenness, PageRank and, the recent EffG centrality (see [32]). This last measure is included in the comparison because it is also a gravitational model based on an effective distance.

Seven real-world networks, with different characteristics and structures, are selected for the calculation of the centrality measures: Jazz, Email, PB, USAir, Physicians, PDZBase and Haggle (or Contact).

4.1. Datasets

The real-world datasets [49] are:

- **Jazz (RELATIONSHIPS)**: a network of collaborations (edges) between jazz musicians and bands (nodes) between 1912 and 1940.
- **Email (COMMUNICATION)**: network representing the exchange of emails (edges) between workers (nodes) of the Rovira i Virgili University in Spain (2003).
- **PB or Political Blogs (SOCIAL)**: hyperlink network between web-blogs on United States politics in 2005, where the weight of edges indicates the political leaning (0.25 means from a liberal or left-wing blog to another left-wing blog, 0.5 means from conservative or right-wing blog to left-wing blog, 0.75 means from left to right, and 1 means from right to right).

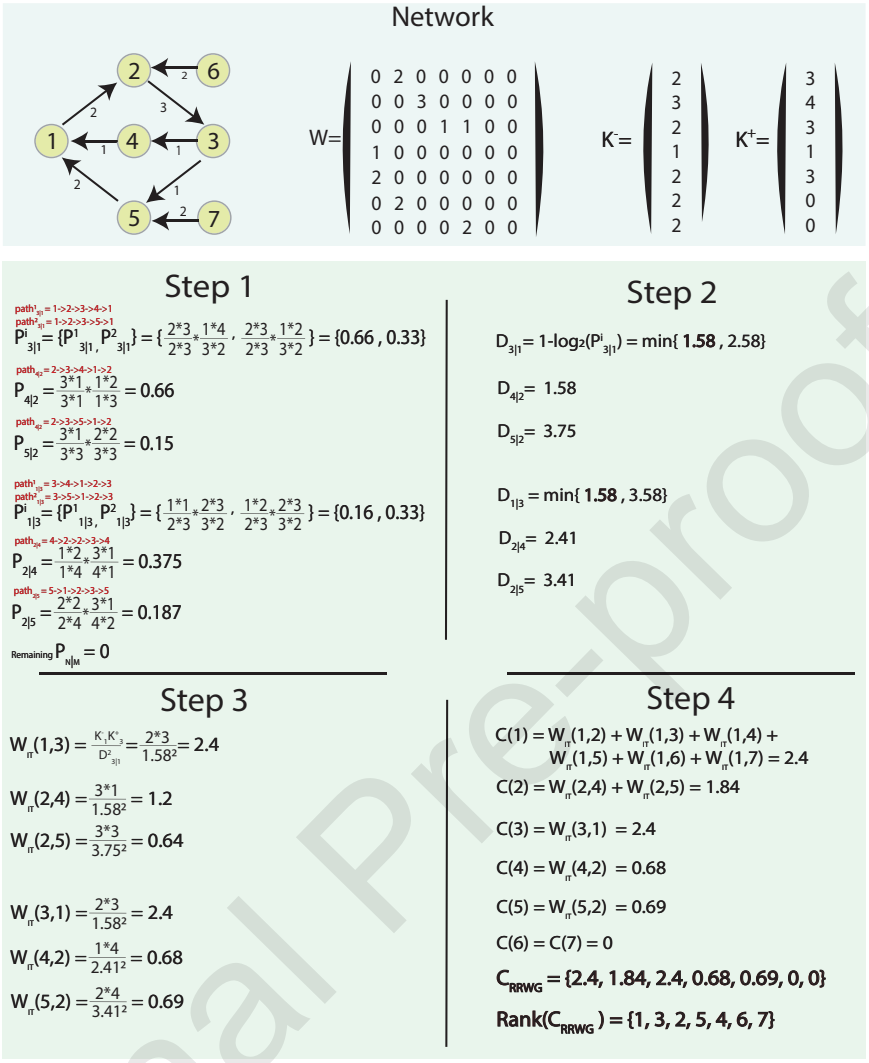


Figure 2: Toy sample of a directed weighted network (top). All possible Return Random Paths are shown in red, and the final results of centrality values and the ranking are highlighted in bold.

- **USAir97 (TRANSPORT)**: a modified version of the original network of United States airport infrastructure, where edges indicate whether there are direct flights between the airports (nodes).
- **Physicians (RELATIONSHIPS)**: network of the trust relationships between the physicians of four cities of the United States in 1966. Nodes are the physicians, and each edge indicates that a destination node trusts or asks for advice from an origin node.
- **PDZBase (SCIENCE)**: a network of PDZ-domain-mediated protein–protein (nodes) binding interactions (edges).
- **Haggle or Contact (CONNECTIVITY)**: a network of human proximities, as measured by user-carried wireless devices. Each person is a node, and each edge indicates when two people were within a certain proximity of each other (measured by a wireless protocol - Bluetooth or Wifi-).
- **EU airlines (TRANSPORT)**: a network of airline routes among European airports. Nodes are the airports and edges are the flight connections between two airports.

- **US Air Traffic (TRANSPORT)**: modified network with the number of commercial flights during a year between two airports in the United States. Nodes are the airports, edges are the flights and weight is the total number of flights in a year.
- **FAA Routes (TRANSPORT)**: network of air traffic routes of the Federal Aviation Administration. Nodes represent airports, and a directed edge is the preferred route between airports i and j.
- **Open Flights (TRANSPORT)**: network of regularly occurring flights between airports worldwide. Nodes are airports, and the direction of edge (i,j) indicates a regularly occurring commercial flight by a particular airline from airport i to airport j. Weights of the arcs or edges are the frequency of the flights.

To validate the C_{RRWG} centrality measure in different kinds of real applications (transport, communication, social networks, relationships, science or wireless mobile network), a series of heterogeneous datasets are selected. Table 3 shows key information about the topology of these networks. Four of these datasets are highlighted because they are a representation of different kinds of networks: Jazz (undirected unweighted graph), Email (directed unweighted graph), PB (directed weighted graph) and USAir (undirected weighted graph). We use these networks to validate the measure by means of a comparison.

Additionally, we use representative sets of real data, related to air transport, to compare C_{RRWG} with different community-aware measures of centrality.

Networks	n	m	$< k >$	$< t >$	C	r	Directed	Weighted
Jazz	198	2472	27.69	8.7	0.63	0.02	NO	NO
Email	1133	5451	9.62	7.75	0.11	0.07	YES	NO
PB	643	4560	12.81	14.66	0.25	0.2	YES	YES
USAir	332	4252	12.81	6.73	0.74	-0.21	NO	YES
Physicians	241	1098	9.11	3.58	0.25	-0.05	YES	NO
PDZBase	212	2672	24.62	50.97	0.3	0.09	NO	NO
Haggle	274	28244	206.16	2.49	0.56	-0.47	NO	NO
EU Airports	417	2953	14.16	11.86	0.38	-0.05	NO	NO
FAA Routes	1226	2615	2.13	77.62	0.07	0.05	YES	NO
US Air Traffic	2278	6390340	2805.24	41.03	0.92	0.03	NO	YES
Open Flights	3214	66771	20.78	51.53	0.34	0	YES	YES

Table 3: Dataset topology information of seven real networks used, where n is the number of nodes, m the number of edges, $< k >$ is the average degree, $< t >$ is random walk mixing time, C is the clustering coefficient and r is the assortative coefficient .

4.2. Comparison with Classical Centrality measures

Firstly, we validate the measure by comparing it with classical centrality measures, such as Degree, Closeness, Betweenness and Pagerank Centrality in four representative datasets with different properties, sizes and characteristics. Additionally, we also compare the proposed metric with a similar approach, the Effective distance gravity model (EffG) [32].

4.2.1. Ranking results

In the first experiments, we compare C_{RRWG} with different methods using the Susceptible-Infected model. Thus, if a node is important, from the point of view of centrality, it has a stronger infectivity in the graph. Two parameters used in the SI model are: the probability of node infection $\beta = 0.2$, which controls the scale over which a node exerts influence, and the average number of infected nodes at that time, which, in this case is $t = 20$. Note that the total number of infected nodes plus the total number of susceptible nodes is always equal to the number of nodes.

a) USAir network.

This is an undirected (symmetric adjacency matrix, Figure 3a-top) and weighted network, where the weight of an edge is the frequency of flights between two airports (nodes). Since our approach, the C_{RRWG} centrality, and the EffG method are both on the Gravity model, a more specific and detailed comparison is made (see Figure 3a-bottom). In this figure, it can be observed that the C_{RRWG} centrality increases the significance of airports belonging to important hubs of connections with other airports, in terms of flight frequencies. In Table 4, we can see that all rankings are similar in the top 10, and equal in the top 3 comparing C_{RRWG} with EffG (*Chicago O'hare International, Dallas/Fort Worth International, and The William B. Hartsfield Atlanta*). However, it is worth noting the importance in the C_{RRWG} ranking of an airport, node 248 (*Los Angeles International*), which only appears in the top 10 of the SI ranking. This difference evidences that C_{RRWG} helps to increase the importance of a node in dense clusters/communities.

#Rank	C_{RRWG}	**EffG**	**DG**	**CC**	**BT**	**PG**	**SI**
1	118	118	118	118	118	118	118
2	261	261	261	261	8	261	261
3	255	255	255	67	261	182	201
4	67	182	152	255	201	152	166
5	166	152	182	201	47	255	182
6	182	166	230	182	182	230	255
7	248	230	166	47	255	166	47
8	201	67	67	166	152	201	67
9	152	112	112	248	313	67	248
10	147	147	201	112	13	8	112

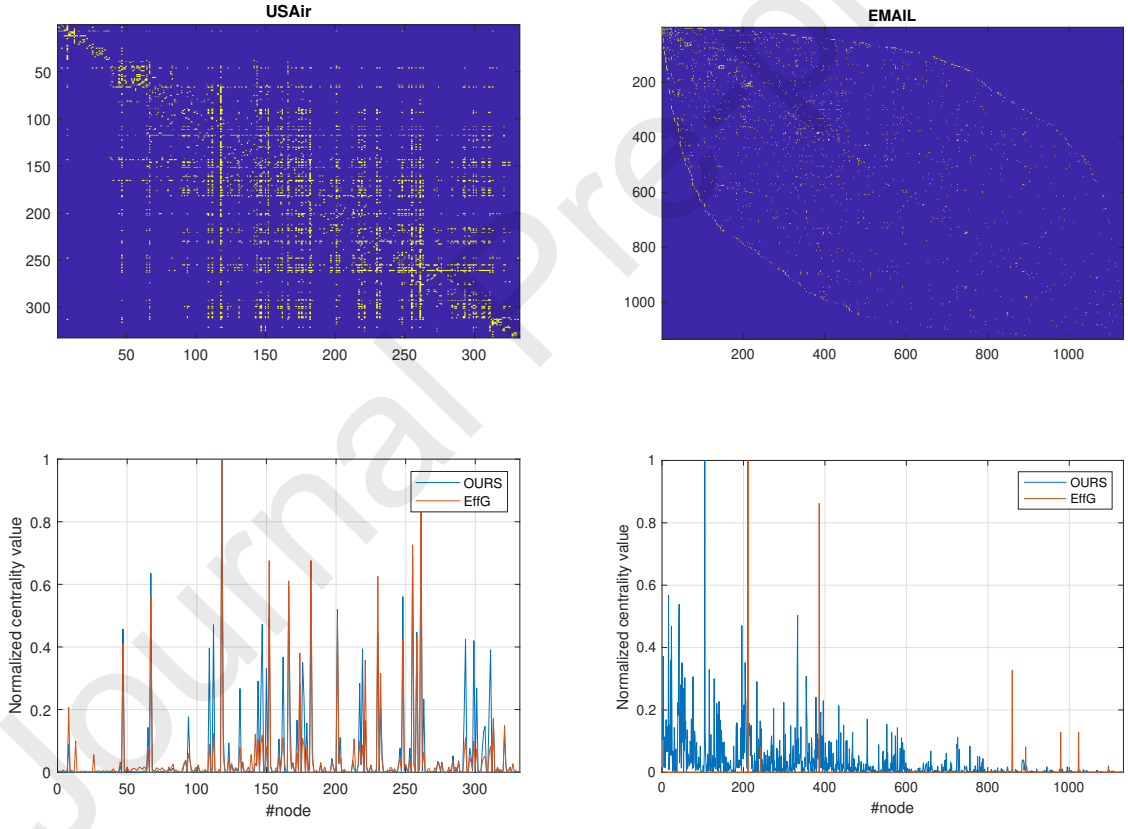
Table 4: Top 10 nodes in USAir network using several methods: C_{RRWG}, EffG, DG, BT, CC, PG and SI.

b) Email network.

This represents the mailing connectivity between workers of a university. This network is directed (asymmetric, but almost symmetric because a person usually answers the received emails), and very sparse. In Fig. 3b-bottom, we can see that the C_{RRWG} gives more importance to first nodes in contrast to EffG centrality because it does not highlight the workers with high degree, closeness, betweenness or Pagerank, focusing on dense clusters or highly connected groups of workers. This effect can be appreciated in Table 5, which shows how our proposal has more similarities with the classic centralities than with the EffG method (nodes with higher ID than in C_{RRWG} ranking).

#Rank	C_{RRWG}	EffG	DG	CC	BT	PG	SI
1	105	211	105	333	333	105	333
2	16	386	333	23	105	23	23
3	42	860	16	105	23	333	42
4	333	979	23	42	578	41	76
5	196	1023	42	41	76	42	41
6	23	242	41	76	233	233	42
7	41	893	196	233	135	16	135
8	3	1096	233	52	41	355	378
9	21	1107	21	135	355	21	355
10	204	105	76	378	42	24	3

Table 5: Top 10 nodes in Email network using several methods: C_{RRWG}, EffG, DG, BT, CC, PG and SI.



(a) USAir (b) Email

Figure 3: Comparison of rankings in USAir and Email networks. Top: adjacency matrices. Bottom: comparison of rankings between the C_{RRWG} centrality and the EffG.

c) PB network.

Political Blogs is a network of hyperlinks between weblogs on United States politics in 2005, where the weight of edges indicates the political leaning. This has two main clusters or communities (liberal blogs and conservative). In Fig. 4, we can see that the densest cluster (left) has more interconnected nodes (blogs) and their centrality values are higher in the C_{RRWG} centrality than in the EffG model. This effect can be observed in the rankings of Table 6, where 80% of the top 10 centrality values of the C_{RRWG} metric belong to the first cluster (the densest community), in contrast to the other methods, which have a minus rate of nodes in the ranking belonging to this cluster.

#Rank	C_{RRWG}	EffG	DG	CC	BT	PG	SI
1	22	318	318	415	318	318	415
2	318	415	391	318	415	391	318
3	224	224	22	439	263	22	32
4	32	32	415	629	211	383	59
5	46	22	32	211	391	439	224
6	59	570	46	570	439	415	383
7	263	46	224	279	570	275	115
8	115	275	439	263	196	410	275
9	65	383	383	383	629	638	211
10	391	115	115	490	54	407	46

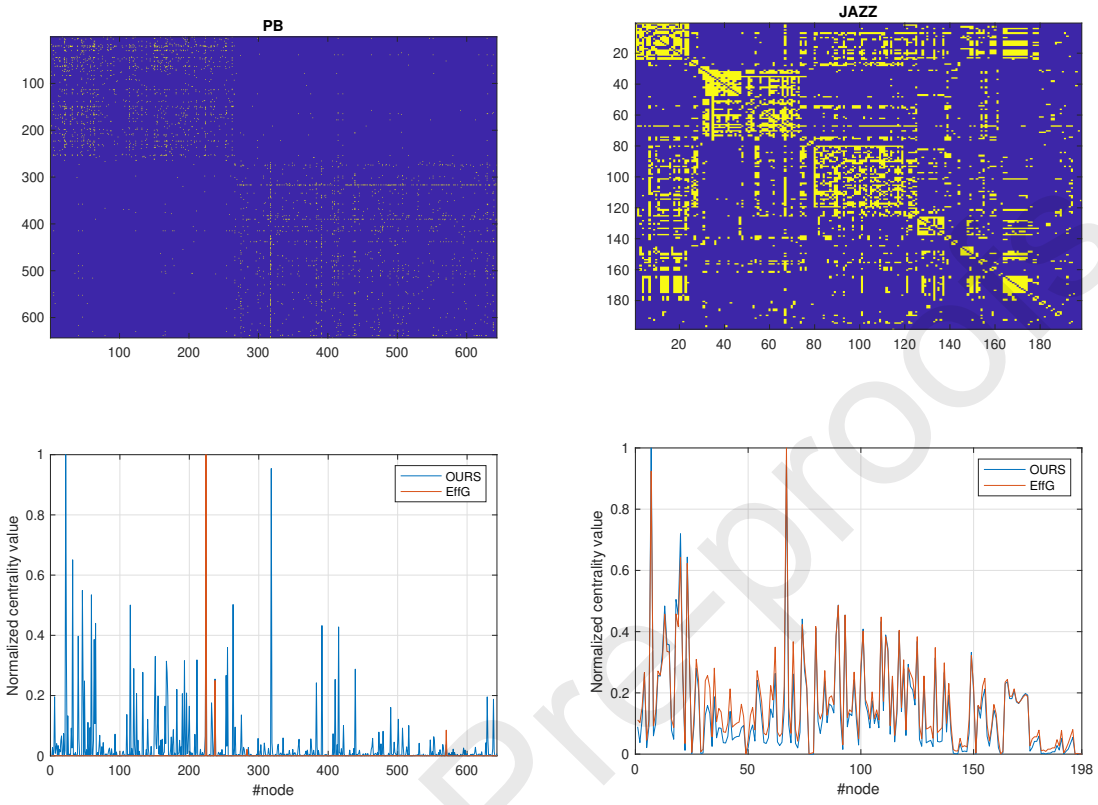
Table 6: Top 10 nodes in PB network using several methods: C_{RRWG}, EffG, DG, BT, CC, PG and SI.

d) Jazz network.

This is a symmetric network (undirected), where there are different clusters. In Fig. 4 we can see three large clusters (around the first 120 nodes) and the remaining network have been formed by small groups. In this case (undirected unweighted graph), the top 10 ranking of the EffG and C_{RRWG} are similar. An important change in the comparison is the most influential node: in C_{RRWG} measure, this node belongs to the first cluster (the densest cluster of the network, and one highly interconnected with nodes of other communities), whereas the EffG indicates that the most important node is 67 belonging to the second cluster (a larger cluster than first) which is the same node as in the other compared rankings (see Table 7).

#Rank	C_{RRWG}	EffG	DG	CC	BT	PG	SI
1	7	67	67	67	67	67	67
2	67	7	7	7	31	7	7
3	20	20	20	23	7	23	23
4	23	23	23	90	70	20	90
5	18	90	90	93	23	70	125
6	90	13	13	20	32	32	74
7	13	93	18	74	47	90	93
8	19	18	93	101	30	93	20
9	93	109	109	125	62	62	46
10	74	74	80	109	93	35	91

Table 7: Top 10 nodes in Jazz network using several methods: C_{RRWG}, EffG, DG, BT, CC, PG and SI.



(a) PB (b) Jazz

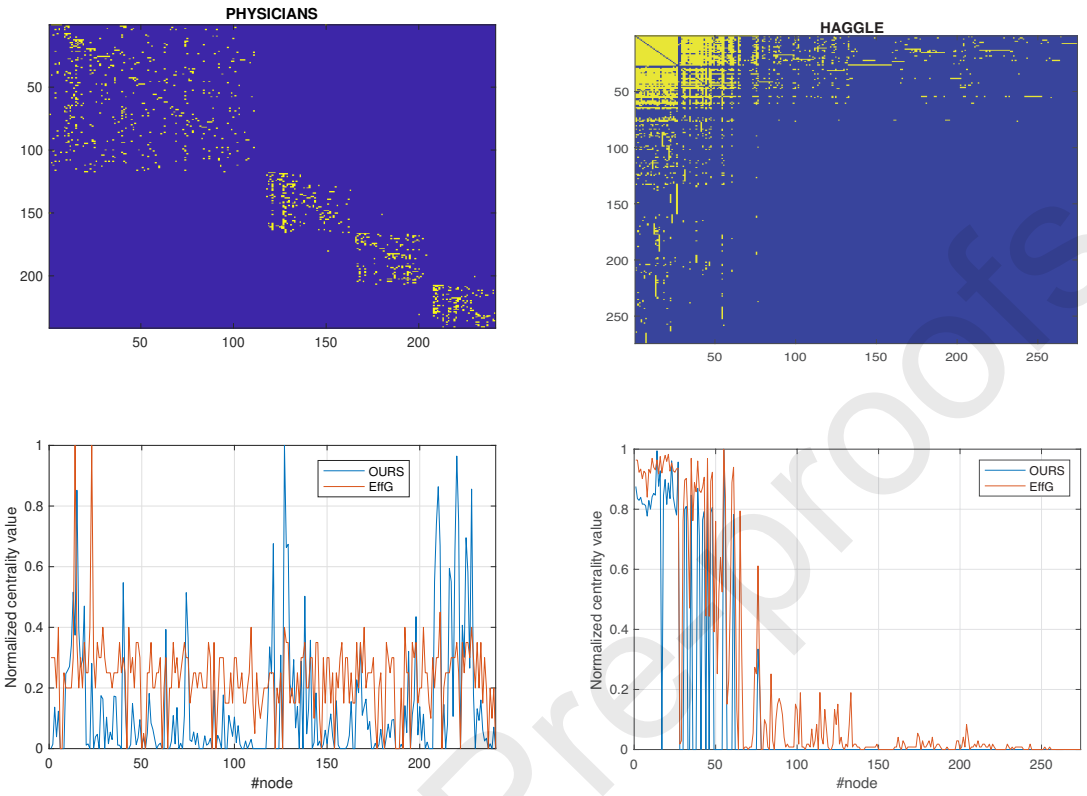
Figure 4: Comparison of rankings in PB and Jazz networks. Top: adjacency matrices. Bottom: comparison of rankings between the C_{RRWG} centrality and the EffG.

e) Physicians, Haggle and PDZbase network.

In this case, we have two particular types of network: Physicians, a network formed by 4 disconnected components or clusters (there are no inter-cluster links), and Haggle, with a single dense cluster with edges to isolated nodes.

In Physicians (Fig. 5a), EffG and the other methods can not differentiate many properties of the network. In contrast, the C_{RRWG} measure has a different ranking (see Table 8), where it focuses on the second and fourth groups (the densest clusters).

The Haggle (Fig. 5b) and PDZbase datasets present a similar behaviour to that of Jazz, with both being undirected unweighted networks with highly similar rankings. However, we can appreciate that the C_{RRWG} centrality considers the isolated nodes as noise because the random path cannot come back in different ways.



(a) Physicians (b) Haggle

Figure 5: Particular networks: Physicians (left) and Haggle (right). Comparison of ranking between EffG and the C_{RRWG} (bottom), showing their adjacency matrices (top).

#Rank	C_{RRWG}	EffG	DG	CC	BT	PG	SI
1	127	15	127	15	15	127	15
2	220	74	15	40	74	15	40
3	210	40	121	74	23	121	12
4	228	11	74	11	11	128	74
5	15	12	128	12	40	74	11
6	221	23	23	23	127	23	69
7	209	127	40	69	12	194	23
8	225	69	11	29	69	40	54
9	121	13	12	54	29	11	13
10	129	10	10	4	10	12	16

Table 8: Top 10 nodes in Physicians network using several methods: C_{RRWG}, EffG, DG, BT, CC, PG and SI.

4.3. Susceptible-Infected (SI) model comparison

The second experiment focuses on the potential estimation of transmission in the network. For that purpose, we apply the SI model [50], which gives us the influence of a node through the transmission (or

infection) ability at a specific time.

At time t, a node i can have two states (susceptible and infection) and during the process (K iterations), the infected nodes can infect the susceptible nodes with a probability β. Finally, we obtain the average number of affected nodes ($F(t)$) at time t.

To establish the configuration of the parameters of the model, we rely on the importance of a node in its cluster or community (same values in β and infection time t). The importance of a node is given by high values of $F(t)$.

To measure the importance of C_{RRWG} with respect to the other compared techniques, we conduct the following process: i) we select the top-M nodes of each ranking for all compared methods, ii) these top-M nodes are used as potentially infected nodes with a β probability, and iii) we calculate the average infected nodes ($F(t)$) for each measure separately, at each different time t.

In the following results, we have used $\beta = 0.2$, $F(t)$ is denoted in figures as $< N >$, and we compare all methods with two values of M (100 and 25 nodes). The node with the highest $F(I)$ is the most important, and one method is better than another if the curve rises faster.

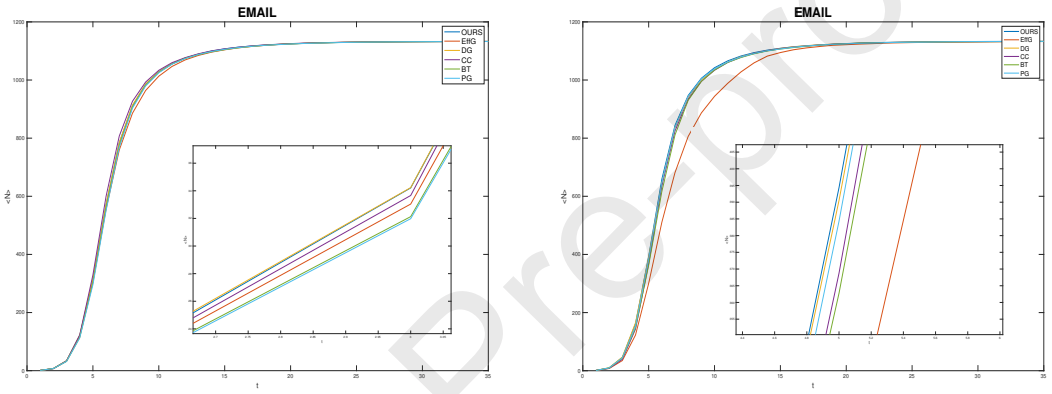


Figure 6: Email - SI model. Comparison of the transmission/infection curves of the top nodes in the SI model (left: top $M = 100$ nodes; right: top $M = 25$ nodes) for all methods. N (y-axis) is the number of infected nodes at time t (x-axis).

In Fig. 6-left, we show the comparison of the curve of infection in the Email network for $M = 100$ top nodes. The C_{RRWG} and DG have similar curves. The least efficient method in this network is EffG, although, in fact, there are no consistent differences. As a complementary experiment (see Fig. 6-right), we focus on the top 25 nodes in each ranking, where C_{RRWG} improves on the results of the remaining methods, and the EffG curve rises more slowly than the others. This is due to the high centrality values of sparse nodes in the EffG (see the top right-hand nodes in Fig. 3b-bottom).

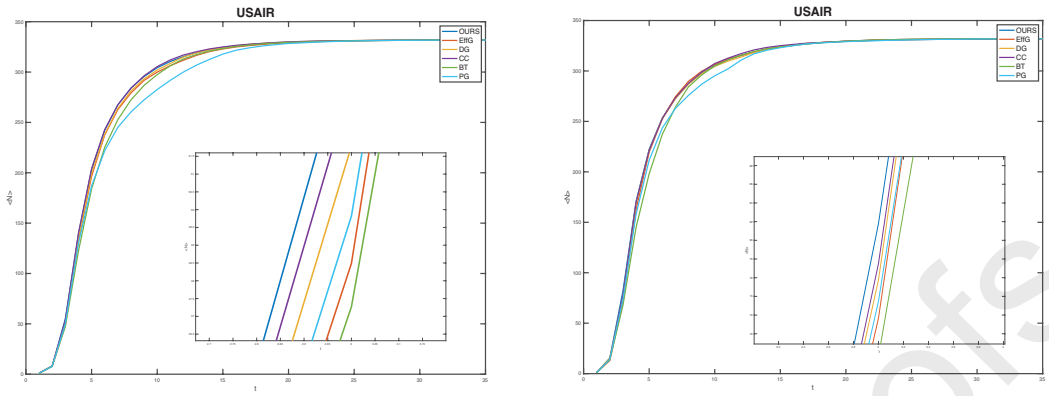


Figure 7: USAir - SI model. Comparison of the transmission/infection curves of the top nodes in the SI model (left: top $M = 100$ nodes; right: top $M = 25$ nodes) for all methods. N (y-axis) is the number of infected nodes at time t (x-axis).

In Fig. 7 for the USAir network, the curves are similar but the proposed method is higher and rises a slightly faster than the others since the top nodes obtained in C_{RRWG} are more important. Closeness works well too, in contrast to PageRank, which is the worst-performing method in this case (this makes sense due to the importance of the closeness in airports networks to link them with frequent flights).

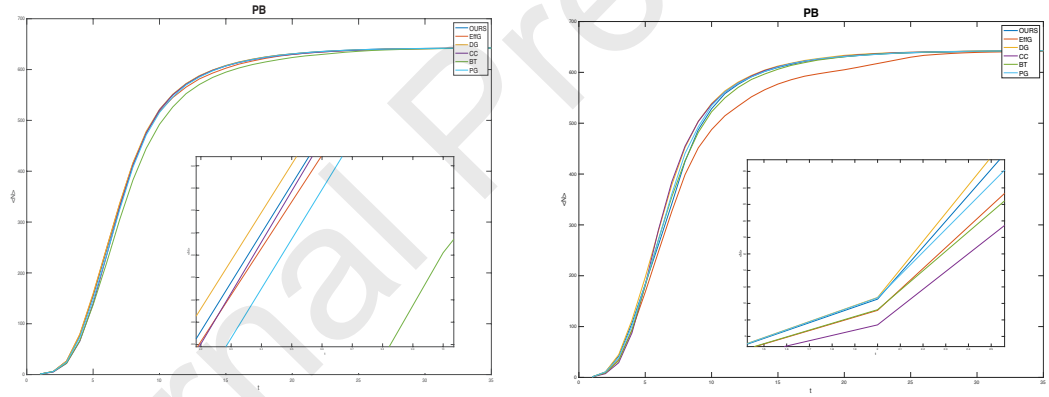


Figure 8: PB - SI model. Comparison of the transmission/infection curves of the top nodes in the SI model (left: top $M = 100$ nodes; right: top $M = 25$ nodes) for all methods. N (y-axis) is the number of infected nodes at time t (x-axis).

In Fig. 8 for the PB network, the curves grow more slowly than in previous networks, because it has few inter-cluster links (links between blogs with different political ideologies). Degree and C_{RRWG} are the fastest curves because they form the densest network of the experimental set. In the case of EffG, the curve is slowing down in Fig. 8-right, where only the top 25 nodes are selected.

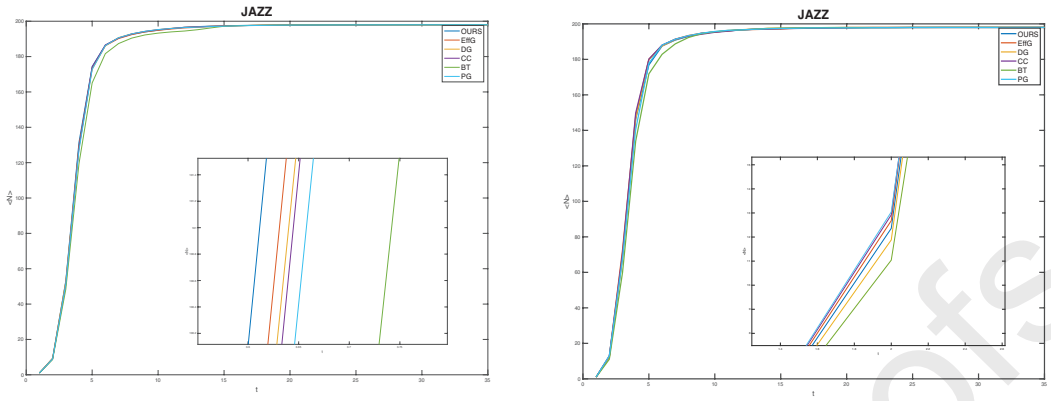


Figure 9: Jazz - SI model. Comparison of the transmission/infection curves of the top nodes in the SI model (left: top $M = 100$ nodes; right: top $M = 25$ nodes) for all methods. N (y-axis) is the number of infected nodes at time t (x-axis).

Regarding the Jazz network, in Fig. 9, the curves grow rapidly because it is a small undirected and unweighted network (198 nodes) with a high density and average degree (27.69), with dense clusters and high inter-cluster connectivity. With these conditions, EffG and C_{RRWG} work better than the others, but all are similar. Nevertheless, the proposed method improves the performance of all measures.

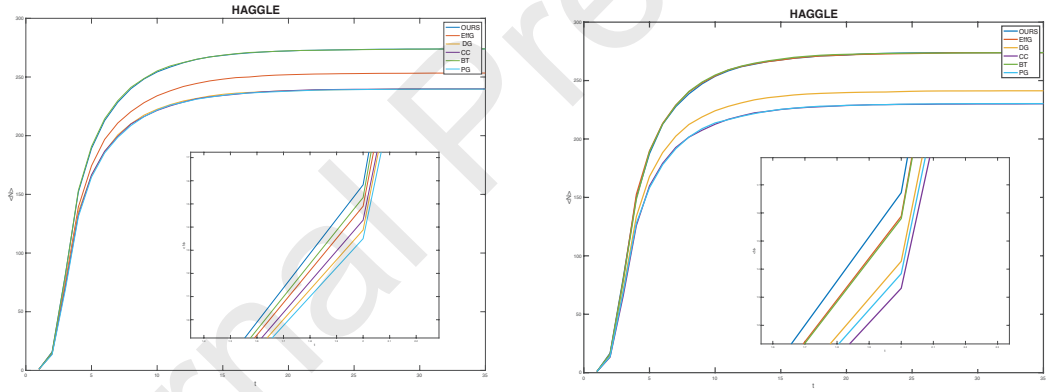


Figure 10: Haggle - SI model. Comparison of the transmission/infection curves of the top nodes in the SI model (left: top $M = 100$ nodes; right: top $M = 25$ nodes) for all methods. N (y-axis) is the number of infected nodes at time t (x-axis).

Finally, we briefly discuss the experiments in the remaining datasets. In Physicians, the infection cannot be propagated to other clusters because there are no inter-cluster edges (four disconnected clusters). The proposed method shows great effectiveness because the density in clusters is high. PDZbase network has similar results to the Jazz network but the Haggle dataset exhibits a different behaviour (see Fig. 10). BT and C_{RRWG} are the highest and rise considerably faster than other methods (PG, DG and CC). This is due to the connectivity concept of this network (wireless devices links), where these measures better capture these network topologies. EffG performance is improved in the reduced case (Fig. 10-right).

Summarising, the curves corresponding to the present proposal respond better in all cases, with different topology networks. This means that the C_{RRWG} measure has the ability to capture different properties of the topology of the network. With the top 100 nodes reduced to the top 25, we demonstrate that the proposed method is robust in terms of the selection of the initial infected nodes, selecting influential ones with greater precision.

4.4. Evaluating infection speed

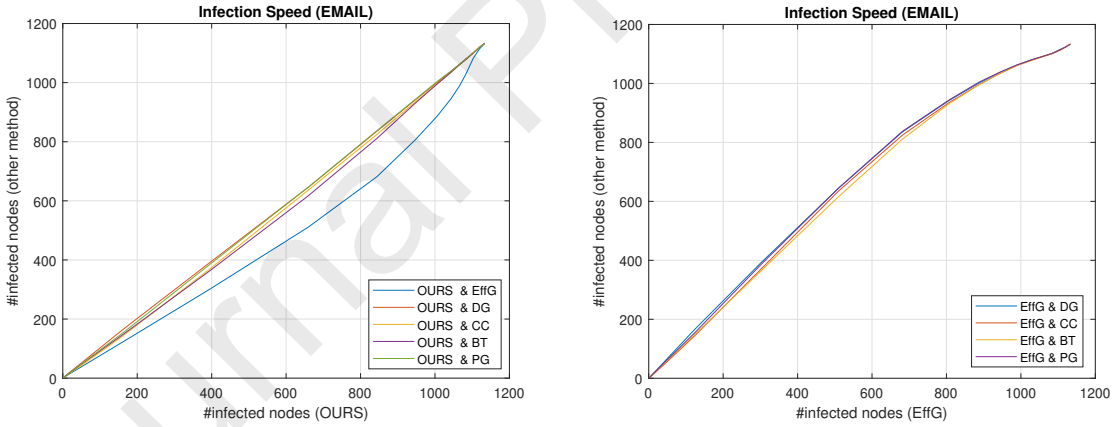
In the third experiment, we evaluate the infection speed through a comparison between correlations. Correlations are first established between C_{RRWG} and the rest of the measures (DC, CC, BT, PG and EffG), and then, correlations are made between the other measure based on the gravitational model (EffG) and the rest of the classic centrality measures (DC, CC, BT and PG). Finally, these correlations are compared.

Three datasets with major differences, Email, USAir and Haggle, are selected with the aim of calculating the correlation of the average number of infected nodes ($F(t)$) of the C_{RRWG} measure (represented in x-axis) and the other centralities (y-axis). This is then compared with the correlation of the EffG metric with the rest of the methods. In the analysis of the graphs, it must be taken into account that if a curve is below the diagonal, the number of infected nodes of the x-axis method is greater than that of the axis method, which means that there are more influential nodes.

Analysing Fig. 11a, we show that the top nodes of the proposed measure infect the network (transmission) more quickly than the EffG model (blue line). Correlations with other methods are more similar, but the curve of the proposed method is slightly better than the other curves. The complementary experiment in Fig. 11b shows that EffG is the worst-performing method in this network because all the curves are above the diagonal.

In the Haggle network in Fig. 12a, the results are similar but the correlation between BT and C_{RRWG} is very high (99%), validating the proposed method as a good estimator of betweenness centrality in dense connected networks. In contrast, EffG cannot satisfactorily estimate the transition role of influential nodes (see Fig. 12b).

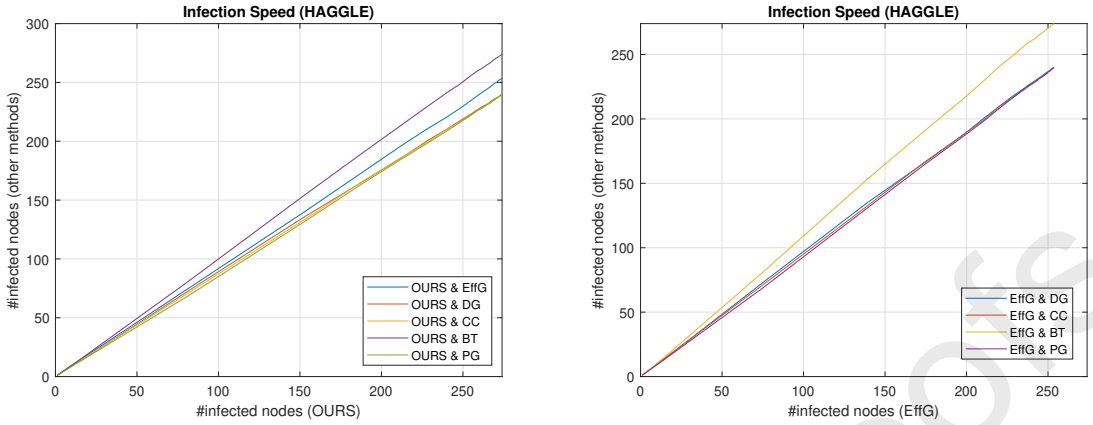
Finally, regarding the USAir network as a typical closeness dataset, the C_{RRWG} method works better than all the methods under comparison (as can be seen in Fig. 13a). In contrast, EffG is not sufficient to improve the performance of the classic closeness centrality (red line in Fig. 13b).



(a) C_{RRWG} method vs other methods

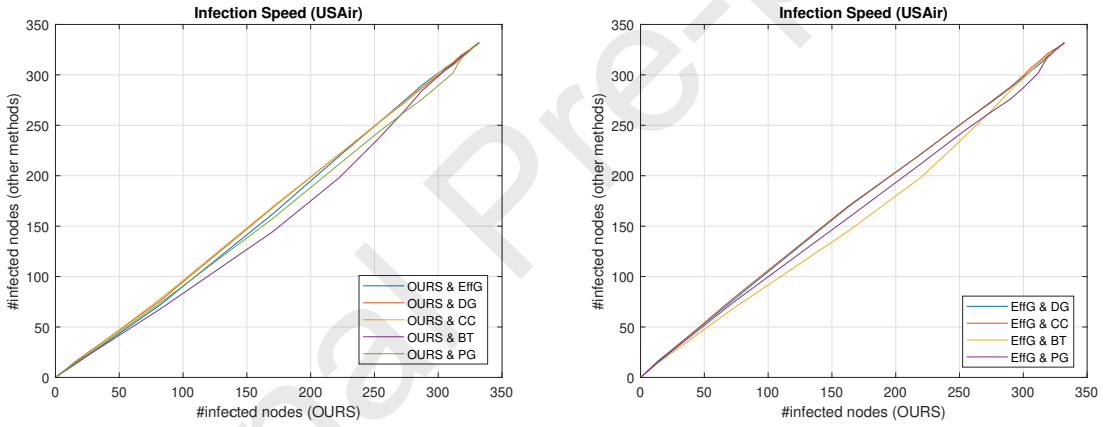
(b) EffG vs other methods

Figure 11: Infection speed (EMAIL): values under the diagonal mean that the method of the x-axis has more infected nodes than the other method.



(a) C_{RRWG} vs other methods (b) EffG vs other methods

Figure 12: Infection speed (HAGGLE): values under the diagonal mean that the method of the x-axis has more infected nodes than the other method.

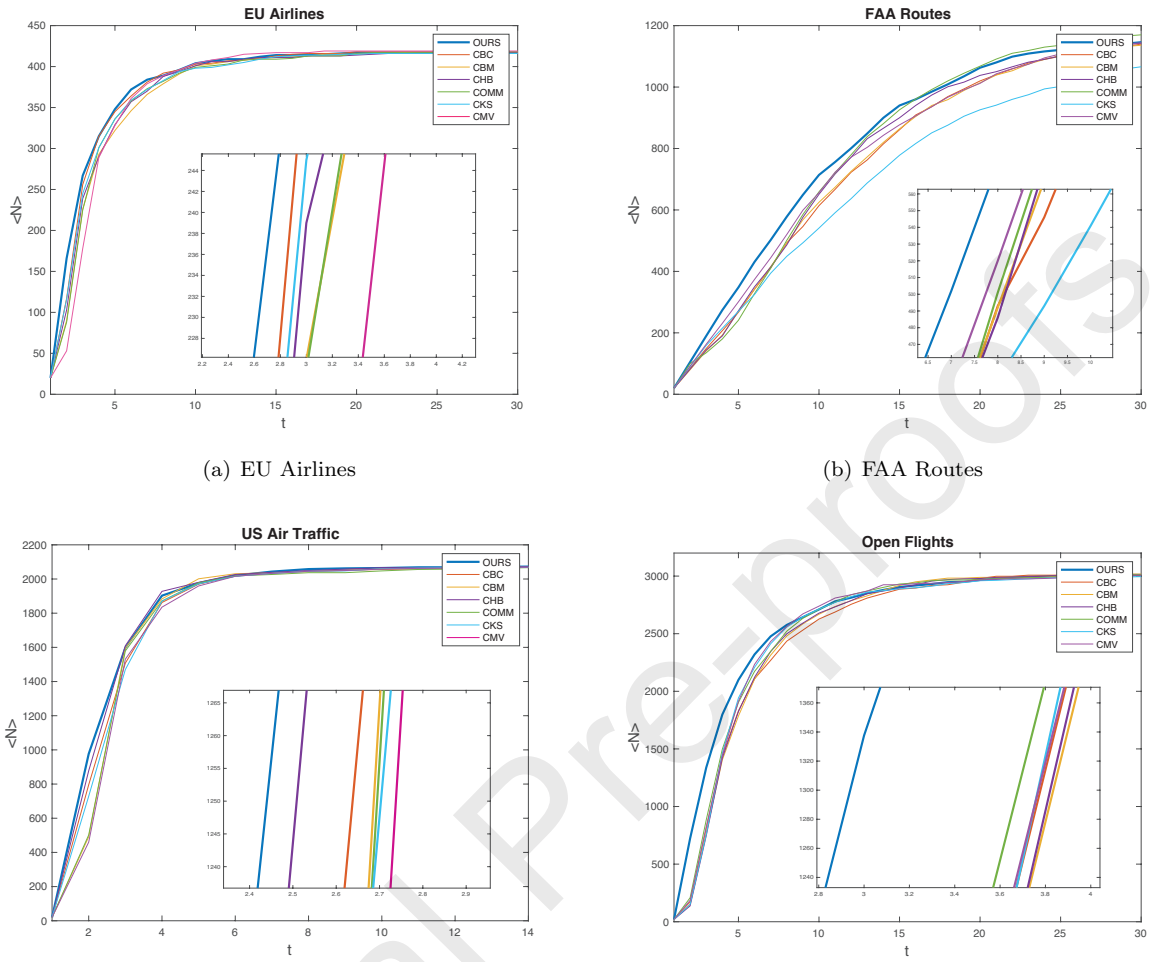


(a) C_{RRWG} vs other methods (b) EffG vs other methods

Figure 13: Infection speed (USAIR): values under the diagonal mean that the method of the x-axis has more infected nodes than the other method.

4.5. Comparison with Community-Aware Centrality measures

Comparing the proposed metric with the classic methods and the EffG approach, we can observe the validity of the method as a useful centrality measure. However, there exist other recent metrics, such as community-aware centrality measures, that exploit local and global properties of the networks and are more realistic alternatives than the classic measures. In order to validate the present proposal, we perform experiments focusing on the comparison with six community-aware centrality measures, using four networks related to air traffic (EU Airlines, FAA Routes, US Air Traffic and Open Flights). These selected datasets have different properties and characteristics (see Table 3), with increasing size to test the scalability of the proposed measure.



(a) EU Airlines

(b) FAA Routes

(c) US Air Traffic

(d) Open Flights

Figure 14: Air Datasets - SI model. Comparison of the transmission/infection curves of the top nodes in the SI model (top M = 25 nodes) for all methods. N (y-axis) is the number of infected nodes at time t (x-axis).

The experiment carried out focuses on estimating the transmission potential in a network using the Susceptible-Infected (SI) model, which shows the influence of a node through its transmission (or infection) capacity at a specific time. At time t, a node can be in two states, susceptible or infected, and during iterations (number of experiments), infected nodes can infect susceptible nodes with probability β. Finally, the average number of nodes affected at time t is obtained. The configuration of the parameters of the model is similar to that presented in Section 4.3. Function $F(t)$ is denoted in the figures as $< N >$, and all the methods are compared with $M = 25$ nodes.

In this comparison, the community-aware centralities used are (see Table 2 for a detailed explanation): Community-based Centrality (CBC), Community-based Mediator (CBM), Community Hub-Bridge (CHB), Comm Centrality (COMM), K-Shell with Community (CKS) and Modularity Vitality (CMV).

Finally, the following process is conducted: i) the top 25 nodes are obtained for each of the seven measures used, ii) these nodes are used as infected nodes with a probability $\beta = 0.2$, in the SI model, and iii) the average infected nodes $F(t)$ are calculated for each measurement separately and for each time t. It is concluded that one measurement is more effective than another depending on the growth rate of the average $F(t)$, that is, one method is more effective than another if the curve rises more quickly.

Figure 14 shows the curves of the comparison of the SI model for seven centrality measures (the proposed measure in dark blue). In the first two networks (a and b), the performance of all measures is similar, although the current proposal is the best-performing. In the biggest datasets (c and d), the improvements are more significant, with the curve of C_{RRWG} growing more rapidly. This is mainly due to their being dense networks, working optimally in this type of network (as can also be seen in the previous experiments). It is important to highlight that each measure works better depending on the network (sparse or dense networks, directed or undirected networks, weighted or unweighted, hubs, bridges, etc.), whereas the proposed method almost always works well, regardless of the type of the network.

As previously discussed, our measure performs optimally on dense networks, but behaves less efficiently on sparse networks, where many end nodes or *leaf nodes* might be unreachable for this approach. This case can be observed in Figure 14-b, where the slope of the curve is less pronounced because the infection speed is limited by the low density of the network. In this situation, C_{RRWG} starts well, better than the remaining measures, but, at time $t = 20$, is outperformed by COMM Centrality, which can reach more nodes than the proposal in this case.

5. Discussion and Conclusions

Recent studies on the identification of influential nodes in real-world networks highlight the importance of considering not only the local and global information of a network but also its dynamic information and the strength of each node in the context of communities. In addition, it would be useful to generalize a measure that covers all types of networks. With the intention of integrating a new centrality measure that brings together these aforementioned objectives, a model for influential node identification, named Return Random Walk Gravity Centrality (C_{RRWG}), is proposed. The measure is based on three approaches: effective distance, gravity model and return random-walk. First, it uses effective distances with the objective of capturing the static and dynamic information of the topology of a network. Secondly, to measure the power of attraction of a node in the network, a gravity model is used. Finally, with the aim of studying and analysing the real strength of each node in a cluster context, taking into account global and local information, return random walks is used. Summarizing, the main aim of the paper is to propose a new centrality measure combining community structure information (local, global, static and dynamic) for all kinds of networks (directed, undirected, weighted and unweighted).

The effectiveness of the proposed centrality (C_{RRWG}) is validated by carrying out a comparison with classic centralities, such as Degree, Closeness, Betweenness or PageRank, and other recent studies based on community structure, such as Community-based Centrality, Community Hub-Bridge, Community-based Mediator, Comm Centrality, K-Shell with Community and Modularity Vitality. In addition, to show the robustness of the centrality presented, it is compared with other centrality measures based on the gravity model with effective distance, as is the case of EffG centrality. The results obtained show how our proposal improves on the performance of all the approaches studied, regardless of the characteristics and properties of the network, for the propagation and detection of the most influential nodes in a network. Moreover, it is especially effective in networks, with the only limitation being the unreachable end nodes or leaf nodes, due to the characteristics of the methodology used. In future research, it would be interesting to conduct a parameterization study of the gravity function to improve the performance in very low density networks.

Acknowledgements. Financial support for this research has been provided under grant PID2020-112827GB-I00 funded by MCIN/AEI/10.13039/501100011033.

References

[1] Shi-Min Cai Chun Yang Ya-Chun Gao, Chuan-Ji Fu and H. Eugene Stanley. Repulsive synchronization in complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(5):053130, 2019.

[2] Rui Ding, Norsidah Ujang, Hussain Bin Hamid, Mohd Shahrudin Abd Manan, Rong Li, Safwan Subhi Mousa Albadareen, Ashkan Nochian, and Jianjun Wu. Application of complex networks theory in urban traffic network researches. *Networks and Spatial Economics*, 19(4):1281–1317, 2019.

[3] Peng Gang Sun and Xiaoke Ma. Dominating communities for hierarchical control of complex networks. *Information Sciences*, 414:247–259, 2017.
[4] Adil Imad Eddine Hosni, Kan Li, and Sadique Ahmad. Minimizing rumor influence in multiplex online social networks based on human individual and social behaviors. *Information Sciences*, 512:1458–1480, 2020.
[5] Qiang He, Xingwei Wang, Fubing Mao, Jianhui Lv, Yuliang Cai, Min Huang, and Qingzheng Xu. Caom: A community-based approach to tackle opinion maximization for social networks. *Information Sciences*, 513:252–269, 2020.
[6] Adil Imad Eddine Hosni, Kan Li, and Sadique Ahmad. Analysis of the impact of online social networks addiction on the propagation of rumors. *Physica A: Statistical Mechanics and its Applications*, 542:123456, 2020.
[7] Ruigang Zheng, Weifu Chen, and Guocan Feng. Semi-supervised node classification via adaptive graph smoothing networks. *Pattern Recognition*, 124:108492, 2022.
[8] Alcebiades Dal Col and Fabiano Petronetto. Graph regularization multidimensional projection. *Pattern Recognition*, 129:108690, 2022.
[9] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4):175–308, 2006.
[10] Peng Gang Sun, Yi Ning Quan, Qi Guang Miao, and Juan Chi. Identifying influential genes in protein–protein interaction networks. *Information Sciences*, 454-455:229–241, 2018.
[11] Zhe Li, Tao Ren, Xiaoqi Ma, Simiao Liu, Yixin Zhang, and Tao Zhou. Identifying influential spreaders by gravity model. *Scientific reports*, 9(1):1–7, 2019.
[12] Sambaran Bandyopadhyay, Ramasuri Narayanam, and M Narasimha Murty. A generic axiomatic characterization for measuring influence in social networks. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2606–2611, 2018.
[13] Wang J Zhang Y Yang W Liu Y Wu X, Cao W. A spatial interaction incorporated betweenness centrality measure. *Plos One*, 17(5):e0268203, 2022.
[14] Xu-Hua Yang, Zhen Xiong, Fangnan Ma, Xiaoze Chen, Zhongyuan Ruan, Peng Jiang, and Xinli Xu. Identifying influential spreaders in complex networks based on network embedding and node local centrality. *Physica A: Statistical Mechanics and its Applications*, 573:125971, 2021.
[15] Phillip Bonacich and Paulette Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23(3):191–201, 2001.
[16] Lihong Han, Qingguo Zhou, Jianxin Tang, Xuhui Yang, and Hengjun Huang. Identifying top-k influential nodes based on discrete particle swarm optimization with local neighborhood degree centrality. *IEEE Access*, 9:21345–21356, 2021.
[17] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.
[18] Xuan Yang and Fuyuan Xiao. An improved gravity model to identify influential nodes in complex networks based on k-shell method. *Knowledge-Based Systems*, 227:107198, 2021.
[19] M.E. J. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54, 2005.
[20] Linyuan Lü, Duanbing Chen, Xiao-Long Ren, Qian-Ming Zhang, Yi-Cheng Zhang, and Tao Zhou. Vital nodes identification in complex networks. *Physics Reports*, 650:1–63, 2016. Vital nodes identification in complex networks.
[21] Roosevelt Sardinha, Aline Paes, and Gerson Zaverucha. Revising the structure of bayesian network classifiers in the presence of missing data. *Information Sciences*, 439-440:108–124, 2018.
[22] Linyuan Lü, Duanbing Chen, Xiao-Long Ren, Qian-Ming Zhang, Yi-Cheng Zhang, and Tao Zhou. Vital nodes identification in complex networks. *Physics Reports*, 650:1–63, 2016.
[23] Taras Agryzkov, Leandro Tortosa, and Jose F. Vicent. A variant of the current flow betweenness centrality and its application in urban networks. *Applied Mathematics and Computation*, 347:600–615, 2019.
[24] Manuel Curado, Rocio Rodriguez, Leandro Tortosa, and Jose F Vicent. A new centrality measure in dense networks based on two-way random walk betweenness. *Applied Mathematics and Computation*, 412:126560, 2022.
[25] Manuel Curado, Rocio Rodriguez, Fernando Terroso-Saenz, Leandro Tortosa, and Jose F Vicent. A centrality model for directed graphs based on the two-way-random path and associated indices for characterizing the nodes. *Journal of Computational Science*, 63:101819, 2022.
[26] Zhuo-Ming Ren, An Zeng, Duan-Bing Chen, Hao Liao, and Jian-Guo Liu. Iterative resource allocation for ranking spreaders in complex networks. *EPL (Europhysics Letters)*, 106(4):48005, may 2014.
[27] Liguo Fei, Qi Zhang, and Yong Deng. Identifying influential nodes in complex networks based on the inverse-square law. *Physica A: Statistical Mechanics and its Applications*, 512:1044–1059, 2018.
[28] Chungu Guo, Liangwei Yang, Xiao Chen, Duanbing Chen, Hui Gao, and Jing Ma. Influential nodes identification in complex networks via information entropy. *Entropy*, 22(2), 2020.
[29] Yue Ma, Zhulou Cao, and Xingqin Qi. Quasi-laplacian centrality: A new vertex centrality measurement based on quasi-laplacian energy of networks. *Physica A: Statistical Mechanics and its Applications*, 527:121130, 2019.
[30] Lien-Fa Lin and Yung-Ming Li. An efficient approach to identify social disseminators for timely information diffusion. *Information Sciences*, 544:78–96, 2021.
[31] Dirk Brockmann and Dirk Helbing. The hidden geometry of complex, network-driven contagion phenomena. *Science*, 342(6164):1337–1342, 2013.
[32] Qiuyan Shang, Yong Deng, and Kang Hao Cheong. Identifying influential nodes in complex networks: Effective distance gravity model. *Information Sciences*, 577:162–179, 2021.
[33] Hocine Cherifi, Gergely Palla, Boleslaw K Szymanski, and Xiaoyan Lu. On community structure in complex networks: challenges and opportunities. *Applied Network Science*, 4(1):1–35, 2019.
[34] Doina Bucur. Top influencers can be identified universally by combining classical centralities. *Scientific reports*, 10(1):1–14,

2020.

[35] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Comparative evaluation of community-aware centrality measures. *Quality & Quantity*, pages 1–30, 2022.

[36] Michael Kitromilidis and Tim S. Evans. Community detection with metadata in a network of biographies of western art painters. 2018.

[37] Zhiying Zhao, Xiaofan Wang, Wei Zhang, and Zhiliang Zhu. A community-based approach to identifying influential spreaders. *Entropy*, 17(4):2228–2252, 2015.

[38] Muluneh Mekonnen Tulu, Ronghui Hou, and Talha Younas. Identifying influential nodes based on community structure to speed up the dissemination of information in complex network. *IEEE Access*, 6:7390–7401, 2018.

[39] Shi-Long Luo, Kai Gong, and Li Kang. Identifying influential spreaders of epidemics on community networks. *arXiv preprint arXiv:1601.07700*, 2016.

[40] Naveen Gupta, Anurag Singh, and Hocine Cherifi. Centrality measures for networks with community structure. *Physica A: Statistical Mechanics and its Applications*, 452:46–59, 2016.

[41] Zakariya Ghalmane, Mohammed El Hassouni, and Hocine Cherifi. Immunization of networks with non-overlapping community structure. *Social Network Analysis and Mining*, 9(1):1–22, 2019.

[42] Thomas Magelinski, Mihovil Bartulovic, and Kathleen M Carley. Measuring node contribution to community structure with modularity vitality. *IEEE Transactions on Network Science and Engineering*, 8(1):707–723, 2021.

[43] Shenghao Liu, Bang Wang, Laurence T. Yang, and Philip S. Yu. Hnf: Hybrid neural filtering based on centrality-aware random walk for personalized recommendation. *IEEE Transactions on Network Science and Engineering*, 9(3):1056–1066, 2022.

[44] Pasquale De Meo, Mark Levene, Fabrizio Messina, and Alessandro Provetti. A general centrality framework-based on node navigability. *IEEE Transactions on Knowledge and Data Engineering*, 32(11):2088–2100, 2020.

[45] Francesca Arrigo, Peter Grindrod, Desmond J Higham, and Vanni Noferini. Non-backtracking walk centrality for directed networks. *Journal of Complex Networks*, 6(1):54–78, 2017.

[46] Manuel Curado. Return random walks for link prediction. *Information Sciences*, 510:99–107, 2020.

[47] Zhe Li, Tao Ren, Xiaoqi Ma, Simiao Liu, Yixin Zhang, and Tao Zhou. Identifying influential spreaders by gravity model. *Scientific Reports*, 9(1):8387, 2019.

[48] Dirk Brockmann and Dirk Helbing. The hidden geometry of complex, network-driven contagion phenomena. *science*, 342(6164):1337–1342, 2013.

[49] Netzschleuder network catalogue, repository and centrifuge. https://networks.skewed.de, 2022.

[50] Tao Wen and Yong Deng. Identification of influencers in complex networks by local information dimensionality. *Information Sciences*, 512:549–562, 2020.