Les Cahiers du GERAD ISSN: 0711–2440

The extended Jaccard distance in complex networks

E. Camby,
G. Caporossi

G–2017–77

September 2017

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

The extended Jaccard distance in complex networks

Eglantine Camby [a, b]

Gilles Caporossi [b]

[a] *Université libre de Bruxelles, Bruxelles, Belgium*

[b] *GERAD & HEC Montréal, Montréal (Québec), Canada, H3T 2A7*

ecamby@ulb.ac.be
gilles.caporossi@hec.ca

September 2017

Les Cahiers du GERAD

G–2017–77

Abstract: Distance measures play an important role in data analysis, mainly for clustering purpose, but also for data representation (for instance using multidimensional scaling) or for prediction (e.g., k-nearest neighbors). If the concept is also well defined in networks, it turns out that the distance measures are either difficult to compute or are not precise enough for most analysis purpose. Furthermore, the concept of distance and its measure should be adapted depending on the application area since it does not have the same meaning in a social network, a telecommunication network or a molecule. In this paper, we propose a new distance measure that is based upon the well known Jaccard distance, but does not have the limitation that all pairs of nodes at geodesic distance strictly greater than 2 automatically have 1 as the Jaccard distance. The new distance measure is defined and analyzed according to its possible applications and in terms of computational complexity.

Keywords: Distance, link prediction, community detection, complex networks

Acknowledgments: This work was partially supported by a post-doc grant "Bourse d'Excellence WBI.WORLD" from Fédération Wallonie-Bruxelles (Belgium) and also by NSERC and by Foundation HEC Montréal (Canada).

1 Introduction

We work on complex network, a mathematical object comprising a set of nodes together with a set of arcs (links between nodes). Let n denote the number of nodes in the network and m its number of arcs.

The most used distance measure in networks is certainly the geodesic distance d_{ij} [3] which counts the minimum number of arcs in a path joining the two considered nodes i and j. If this distance could be very efficiently computed using Dijkstra's algorithm, it only uses few of the information available. Indeed, as soon as the shortest path's length is not altered, the remaining of the network does not matter. This distance is also degenerated in the sense that lots of pairs of nodes have the same geodesic distance. For instance, adjacent nodes are always at distance 1.

To use some more precise information, one could use the resistance distance [9], or the random walk distance [1] which is based upon the very same process [11, 13]. Unfortunately, if the latter distances are more accurate as they do not involve only the length of the shortest path, but also the number of alternative paths of various length, they are difficult to compute. Namely, the Laplacian (of Kirchhoff) pseudo inverse must be computed [6], which is time and memory consuming [12].

A fast alternative to the resistance distance that is commonly used in the case of sparse data, and specifically in text mining, is the Jaccard distance [7, 8]. Initially built for bipartite graphs, the Jaccard distance is based upon the proportion of neighbors of each node that are common to both of them. This definition provides a more precise measure of the Jaccard distance between nodes and gives a value between 0 and 1. Unfortunately, as soon as two nodes do not share any neighbor, the Jaccard distance is always 1, which results in a lack of information.

In this paper, we propose a new distance measure that is based upon the geodesic and the Jaccard distances. Its computation complexity is similar to that of the geodesic distance while it yields the discriminating capability of the Jaccard distance. As such, it seems a good alternative to both the Jaccard and the geodesic distances in the context of complex network analysis.

2 The Jaccard index and the Jaccard distance

The Jaccard distance is directly derived from the Jaccard index which is a measure that aims at quantifying the similarity of objects described by their binary attributes. Let us consider two objects i and j respectively described by the sets of attributes a_i and a_j. The Jaccard index is defined as :

$$J_{ij} = \frac{|a_i \cap a_j|}{|a_i \cup a_j|}.$$

The value of the Jaccard index is always between 0 and 1, and a larger value implies a larger similarity. A distance measure may therefore be defined as

$$DJ_{ij} = 1 - J_{ij}.$$

Since the description of objects by their binary attributes may also be modeled as a bipartite graph where nodes may represent either an object or an attribute and arcs represent the belonging relation, it is possible to translate the definition of the Jaccard index as follows :

$$J_{ij} = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|}, \tag{1}$$

where $N(i)$ denotes the set of nodes adjacent to i, the neighbors of i, and i itself. As previously, the Jaccard distance for bipartite graphs is expressed as :

$$JD_{ij} = 1 - J_{ij}$$

and is always a value between 0 and 1.

3 The Extended Jaccard distance

The definition from Equation (1) may be applied in the case of non necessarily bipartite graphs. We notice that $JD_{ij} = 1$ for all i and j such that $d_{ij} > 2$, which indicates an absence of information for pairs of nodes that are not close to each other. To better measure the distance between pairs of nodes that are at a larger geodesic distance, the definition of the neighborhood $N(i)$ could be modified by introducing a parameter k as follows :

$$N^k(i) = \{j \mid d_{ij} \leqslant k\},$$

i.e. $N^k(i)$ is the set of nodes at geodesic distance at most k from the node i. By varying the parameter k, a set of Jaccard indices may be computed as:

$$J_{ij}^k = \frac{|N^k(i) \cap N^k(j)|}{|N^k(i) \cup N^k(j)|},$$

and the corresponding Jaccard distances are:

$$JD_{ij}^k = 1 - J_{ij}^k.$$

Depending on the value k, the sensitivity of the measures will depend on the geodesic distance between the considered nodes. Indeed, if k is large enough, the index $J_{ij}^k = 1$ and the corresponding Jaccard distance is 0. On the other hand, if k is too small, $J_{ij}^k = 0$ for pairs of nodes at a large geodesic distance from each other. Since it is not convenient to adjust the parameter k to each pair of nodes, and since a distance measure should be meaningful and comparable in all cases, we propose to combine those measures into a single one by summation. We thus define the Extended Jaccard Distance as follows :

$$EJD_{ij} = \sum_{k=0}^{n} JD_{ij}^k.$$

The summation starts with $k = 0$ to ensure that distinct nodes will have a non-zero distance, actually at least 1. Also, if two nodes are close enough, after some value k, the contribution of $1 - J_{ij}^k$ will be 0. On the other hand, if two nodes are far enough, the value $1 - J_{ij}^k$ will be 1 for small values of k, which will contribute to increase this new distance.

4 Correctness of the Extended Jaccard distance

Theorem 1 *If N is the set of nodes in the network, then the function*

$$EJD_{\cdot\cdot} : N \times N \to \mathbb{R}^+ : (i, j) \mapsto EJD_{ij}$$

defined by

$$EJD_{ij} = \sum_{k=0}^{n} JD_{ij}^k$$

is a distance.

Proof. This proof is divided into the four following lemmas. □

Lemma 1 *(Positivity)*

$$EJD_{ij} \geqslant 0 \quad \forall i, j.$$

Proof. Since EJD_{ij} is the sum of the terms JD_{ij}^k for $k = 0 \dots n$, and

$$0 \leqslant J_{ij}^k = \frac{|N^k(i) \cap N^k(j)|}{|N^k(i) \cup N^k(j)|} \leqslant 1,$$

we have $0 \leqslant JD_{ij}^k = 1 - J_{ij}^k \leqslant 1$. The result follows. □

Lemma 2 *(Separability)*

$$EJD_{ij} = 0 \Leftrightarrow i = j.$$

Proof. First, note that $JD^k ii = 0$ for all k, therefore $EJD_{ii} = 0$. Second, since $N^0(i) = \{i\}$, it follows that $JD^0_{ij} = 1$ for all distinct i and j. As EJD_{ij} is a sum of non negative terms, it follows that as soon as one of those terms is greater than 0, the summation will also be, which is the case when $i \neq j$. □

Lemma 3 *(Symmetricity)*

$$EJD_{ij} = EJD_{ji} \quad \forall i, j.$$

Proof. This is due to the symmetry of the operators of union and intersection between sets of nodes, used in the definition of EJD_{ij}. □

Lemma 4 *(Triangle inequality)*

$$EJD_{ij} + EJD_{jl} \geqslant EJD_{il} \quad \forall i, j, l.$$

Proof. To prove the triangle inequality, it is sufficient to prove that for all i, j, l, and for all k,

$$JD^k_{ij} + JD^k_{jl} \geqslant JD^k_{il}.$$

Moreover, the last inequality is a particular case from the proof of the triangle inequality for the Jaccard distance [10]. □

5 Algorithmic issues

There are various ways to compute the Extended Jaccard distance and depending on the implementation, the computational complexity may vary as well as the memory requirement.

The easiest way to proceed consists in building the geodesic distance matrix first, and compute the Extended Jaccard distance by scanning the rows corresponding to the nodes i and j. The computational complexity is then $O(n^3)$. Unfortunately, this implementation requires to store an $n \times n$ matrix. In the case of complex networks, computing and storing a whole distance matrix is no more possible.

It is also feasible to compute the Extended Jaccard distance between two given nodes by applying a variant of the Dijkstra's algorithm. The computational complexity of this distance for a pair of nodes is then $O(m)$, the overall computational complexity for all pairs would then be $O(n^2 \times m)$. In the case of complex networks, if the number of arcs is linear in the number of nodes, i.e. if the network is sparse, the second implementation is better than the first one because it does not require an $n \times n$ matrix to be stored. Otherwise the computational complexity of the second one becomes $O(n^4)$, which is worse than the first one.

6 Illustration through a clustering example

Since it is recursively based upon the set of neighbors and the similarity between neighbors of given nodes, the Extended Jaccard distance is well suited for instance for social networks, networks involving relations between concepts or any data mining application that relies on sparse matrices (represented as bipartite graphs).

The Extended Jaccard distance matrix provides an efficient tool for communities detections based upon a hierarchical clustering algorithm [4], such as the single linkage [5] or the complete linkage [2]. If the single linkage algorithm is used, the partitions obtained will be very similar to those obtained using the standard

Jaccard distance since two nodes with a small Jaccard distance are within the same value incremented by slightly more than 1 as the Extended Jaccard distance. However, the situation is different when the complete linkage is used. As soon as two nodes are at geodesic distance at least 2, the Jaccard distance is systematically 1 and a part of the clustering involves ties that are randomly broken, which is not the case using the Extended Jaccard distance.

In order to show the difference between various distances, the complete linkage was applied to the geodesic distance, the resistance distance, the random walk distance, the Jaccard distance and the Extended Jaccard distance computed on the well-known Zachary dataset [14], i.e. on Zachary's Karate network with 34 nodes and 78 arcs. The obtained partitions on 6 clusters are shown in Figure 1, Figure 2, Figure 3, Figure 4 and Figure 5.

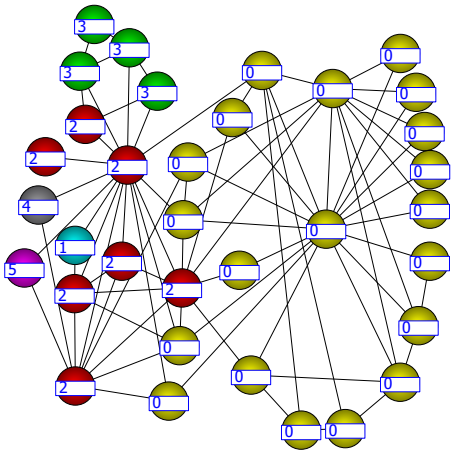

Figure 1: Karate partition on 6 clusters using the geodesic distance.

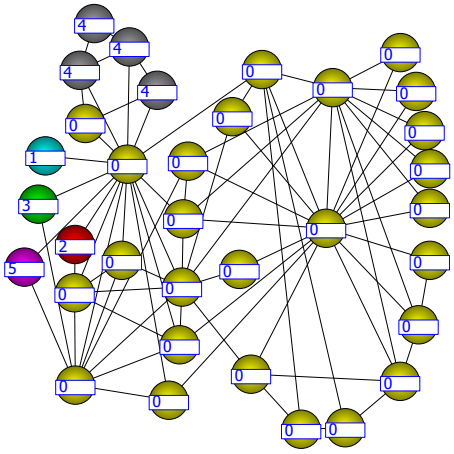

Figure 2: Karate partition on 6 clusters using the resistance distance.

We notice that the path-based distances, i.e. the geodesic, the resistance and the random walk distances, provide comparable results, except maybe the resistance distance for which one large cluster and 4 singletons are found.

On the other hand, the Jaccard-based distances produce another type of result for which a cluster is mainly made of nodes that are not adjacent to any other from that cluster, i.e. the green cluster with number 3 in Figure 4 and Figure 5. This is because these nodes share the same neighbors, and it shows the strong capability of the Jaccard-based distances to predict missing links.

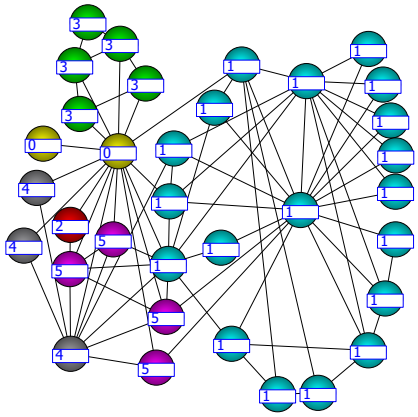

Figure 3: Karate partition on 6 clusters using the random walk distance.

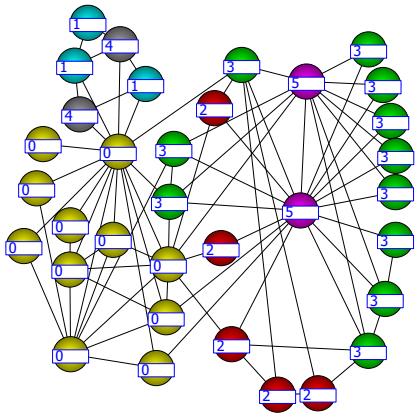

Figure 4: Karate partition on 6 clusters using the Jaccard distance.

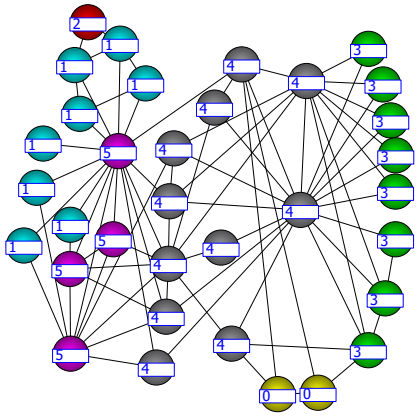

Figure 5: Karate partition on 6 clusters using the Extended Jaccard distance.

Note further that, due to the degeneracy of the standard Jaccard distance, different runs may lead to different partitions. Besides, we can observe that the red cluster with number 2 in Figure 4 is divided with the Extended Jaccard distance in Figure 5. More generally, with the standard Jaccard distance, some clusters are defined arbitrarily whereas the accuracy of the Extended Jaccard distance breaks this arbitrary phenomenon by using the structure of the network.

7 Conclusion and future work

In this paper, we propose an extension of the Jaccard distance to be used in complex networks : the Extended Jaccard distance. This new distance is not restricted to pairs of nodes that have common neighbors, which makes it more convenient as any other distance measures.

The algorithmic complexity for computing the new distance between all pairs of nodes is similar to that of the geodesic, the resistance and the random walk distances, i.e. in $O(n^3)$, when storing an $n \times n$ matrix is possible. If storing such a matrix is not possible, the computation of the Extended Jaccard distance between two given nodes could be achieved in $O(m)$, which is also the case for the geodesic distance, whereas for the resistance and the random walk distances, computing the distance between two given nodes needs to inverse an $n \times n$ matrix, which costs $O(n^3)$.

Beyond these algorithmic issues, the main contribution of this paper is to propose a distance measure for complex networks that is based upon a different paradigm. Unlike the geodesic, resistance or random walk distances, which are defined by the means of paths, the Jaccard and the Extended Jaccard distances are based upon the topological similarities between nodes, which implies that the presence of an arc between two nodes has less impact on their distances than it would have for other distances.

Moreover, the Extended Jaccard distance refines the standard Jaccard distance. Firstly, two distinct nodes have the Extended Jaccard distance at least 1, whereas two nodes sharing same neighbors have 0 as the standard Jaccard distance. Secondly, if the geodesic distance between two given nodes is at least 3 then the standard Jaccard distance is always 1 while the Extended Jaccard distance is more subtle, i.e. a value greater than 2 depending on common nodes which are at geodesic distance at least 3 from the two given nodes.

The first application of the Extended Jaccard distance is thus the detection of spurious arcs or missing ones. Indeed, an arc linking two nodes with a large Extended Jaccard distance indicates a relation between two dissimilar ones, i.e. this arc is potentially spurious. On the reverse, two non-adjacent nodes with a small Extended Jaccard distance should probably be connected. Other applications of the new distance measure are still to be explored.

References

[1] E. Camby and G. Caporossi. Expected distance based on random walks. Journal of Mathematical Chemistry, 2017.

[2] D. Defays. An efficient algorithm for a complete link method. The Computer Journal, 20(4):364–366, 1977.

[3] M.M. Deza and E. Deza. Encyclopedia of distances. In Encyclopedia of Distances, pp. 1–583. Springer, 2009.

[4] J. Friedman, T. Hastie, and R. Tibshirani. The elements of statistical learning, volume 1. Springer series in statistics New York, 2001.

[5] J.C. Gower and G.J.S. Ross. Minimum spanning trees and single linkage cluster analysis. Applied statistics, pp. 54–64, 1969.

[6] I. Gutman and W. Xiao. Generalized inverse of the laplacian matrix and some applications. Bulletin: Classe des sciences mathematiques et natturalles, 129(29):15–23, 2004.

[7] P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. Bull Soc Vaudoise Sci Nat, 37:547–579, 1901.

[8] P. Jaccard. The distribution of the flora in the alpine zone. New phytologist, 11(2):37–50, 1912.

[9] D.J. Klein and M. Randić. Resistance distance. Journal of Mathematical Chemistry, 12(1):81–95, 1993.

[10] M. Levandowsky and D. Winter. Distance between sets. Nature, 234(5323):34–35, 1971.

[11] M.E.J. Newman. A measure of betweenness centrality based on random walks. Social Networks, 27(1):39–54, 2005.

[12] G. Ranjan, Z.L. Zhang, and D. Boley. Incremental computation of pseudo-inverse of laplacian. In International Conference on Combinatorial Optimization and Applications, pp. 729–749. Springer, 2014.

[13] K. Stephenson and M. Zelen. Rethinking centrality: Methods and examples. Social Networks, 11(1):1–37, 1989.

[14] W.W. Zachary. An information flow model for conflict and fission in small groups. Journal of anthropological research, 33(4):452–473, 1977.