A symmetric approach to understand the dynamics of scientific collaborations and knowledge production

Elisa Omodei [1], **Thierry Poibeau** [2], **Jean-Philippe Cointet** [3]

1. *LaTTiCe-CNRS and ISC-PIF*
   elisa.omodei@ens.fr
2. *LaTTiCe-CNRS*
   thierry.poibeau@ens.fr
3. *INRA-SenS and ISC-PIF*
   jean-philippe.cointet@polytechnique.org

ABSTRACT. *How are social and epistemic structures of a research community driving future research dynamics ? We analyzed a very large data set of scientific publications (APS dataset) covering more than 20 years to investigate the social dynamics of collaboration network and the epistemic network of PACS codes co-occurrences network. In particular we show that the epistemic network built through concepts co-occurrences in articles has a non trivial structure characterized by an heterogeneous degree distribution and a high value of modularity, produced by local mechanisms that depend on similarity factors. We empirically show that those local epistemic dynamics are both linked to past social and epistemic structures. Moreover we show that the evolution of the social network depends also on epistemic factors, indicating that the two networks mutually influence their evolution and are thus co-evolving over time.*

RÉSUMÉ. *Comment les structures sociales et épistémiques d'une communauté scientifique contraignent-elles les dynamiques de recherche à venir ? Nous avons analysé un très grand corpus de publications scientifiques (corpus APS) décrivant sur plus de 20 ans de recherche en physique et dont nous avons pu extraire un réseau social de collaboration entre auteurs et un réseau épistémique de cooccurrences entre codes PACS. Nous mettons notamment en évidence que le réseau épistémique a une structure modulaire et une distribution degré hétérogène. La structure en communautés peut sans doute être expliquée par des processus de sélection locaux. Un examen empirique montre que les dynamiques épistémiques locales dépendent aussi bien des structures sociales et épistémiques passées. De plus, nous montrons que l'évolution du réseau social dépend également de facteur épistémiques, ce qui semble indiquer que les deux réseaux co-évoluent l'un avec l'autre.*

KEYWORDS: *socio-semantic networks, co-authorship networks, epistemic networks, knowledge communities, scientific communities dynamics*

Hermès Science Publication – n° y/2013, 1-10

MOTS-CLÉS : réseaux socio-sémantiques, réseaux de collaboration, réseaux épistémique, communauté de savoirs, dynamique des communautés scientifiques

DOI:10.3166/HSP.x.1-10

1. Introduction

The focus of this work is the dynamics of "knowledge communities", defined as systems involving people creating and processing knowledge, connecting concepts in a distributed manner. Examples of such communities are webloggers, wiki contributors and communities of scientists. We will here focus on scientific communities and in particular on the communities of physicists, through the analysis of the APS dataset, that we will introduce in the following section. Many works have analyzed the social dynamics of this community (see in particular Newman's works (Newman, 2001a) (Newman, 2001b)), but to our knowledge no one has carried out a study on its epistemic dynamics, i.e. on the dynamics of the concepts developed and processed by the researchers in the field, with the exception of (Herrera *et al.*, 2010). There have been some studies on the social and epistemic dynamics of some other scientific fields (Roth, 2005), but the concepts considered in that work are a very limited number and so it was not possible to perform a deep analysis of the evolution of the network they form. More recently there has been a new set of works on these topics. For example in (Sun *et al.*, 2013) the authors claim that the evolution of scientific disciplines is purely due to social factors, whereas the goal of our work is to shed new light on the role of concepts in this evolution, and show that the network they form has a non trivial structure that evolves over time in function of social and epistemic factors, in turn possibly influencing also the social dynamics of authors, that could also rely on both the social and the epistemic dimensions.

2. The dataset

The APS dataset[1] contains metadata about over 450000 articles published in the journals Physical Review Letters, Physical Review, and Reviews of Modern Physics from 1893 to 2009. Part of these metadata is the PACS codes characterizing each article. The PACS system is "a hierarchical subject classification scheme designed to classify and categorize the literature of physics and astronomy"[2]. These codes are provided by the authors in order to characterize the content of their papers. About 38% of the articles are assigned three codes, 29% four, 23% two and 10% only one. The codes provided are fourth- and fifth-level codes, representing the most detailed characterization. Each code thus indicates a physical concept that the article addresses, like for example "neutrino interactions" or "solid-liquid transitions". This classification

1. https://publish.aps.org/datasets
2. http://www.aip.org/pacs/

system was introduce in 1970, but it's only from 1985 that the majority of the articles in the dataset are assigned such codes. Since the focus of our work are concepts, we restricted our analysis to the articles published in the years 1985-2009 and having the PACS code information.
Moreover we filtered the dataset by eliminating all the articles with 50 or more authors as recently suggested in (Martin *et al.*, 2013), in order to get rid of the publications in the experimental particle physics domain, which are signed by often hundreds or even thousands of authors. This happens because all the people working in the corresponding consortium are included in the list, even though there was no real direct collaboration between all of them. Therefore the exclusion of these articles avoids introducing topological artefacts in the co-authorship network.

3. Definitions

Following Newman (Newman, 2001a) we define the social network as the weighted undirected graph whose nodes are the authors of the papers in the dataset. Two authors are connected if they co-authored at least one paper, and the weight of the edge between them has value equal to the number of papers they co-authored. We then define the epistemic network as the graph connecting concepts. In our dataset concepts are expressed by the PACS codes. Two PACS codes are connected in the epistemic network if they were both used to classify the same paper at least once, and the weight of the edge between them has value equal to the number of papers in which they appear together. Finally, we define the socio-epistemic network as the graph connecting authors and concepts. The weight of the edge between an author a and a concept c has value equal to the number of articles of which a is an author and c was used to characterize its subject. Every author is thus characterized by a list of concepts he worked on in the frame of his researches, and every concept by a list of authors who used it. A schematic representation of the social, epistemic and socio-epistemic network is given in Figure 1.

4. Results

Firstly we studied the characteristics of the epistemic network. Figure 2 shows that it has both degree and strength distributions that show scale-free properties. In order to further investigate the properties of this network, we compared it with a random network built as follows. The way we constructed the epistemic network through concepts co-occurrences is as if we had firstly built a bipartite network in which the two classes of nodes are the articles and the concepts, and then projected this network into one containing only the concepts and connecting the ones that were connected to the same articles in the original network. Latapy and collaborators showed in (Latapy *et al.*, 2008) that networks built as a projection of a bipartite network present a higher number of edges and a higher clustering coefficient, which are due to the projection process. So in order to create a realistic null model for our network, we first obtained the articles and concepts degree distributions of the real bipartite articles-concepts

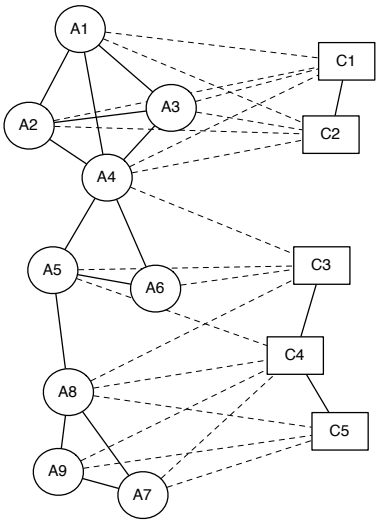

Figure 1. This schematic socio-epistemic network was produced by 4 articles. The first one gathered authors A1, A2, A3, A4 along with PACS codes C1 and C2, the second one A4, A5, A6 along with concept C3. A5 and A8 then published an article using codes C3 and C4. Last, A7, A8 and A9 co-authored an article about C4 and C5.

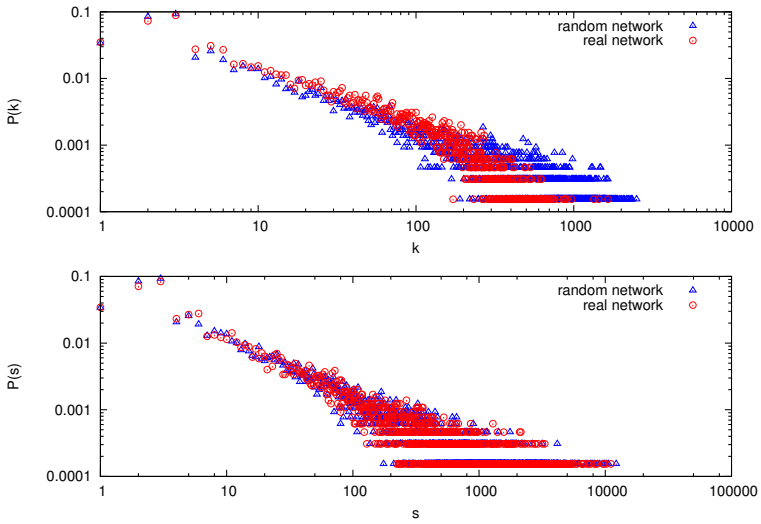

Figure 2. Real (red circles) and random (blue triangles) epistemic network distributions of degrees (up) and strengths (bottom)

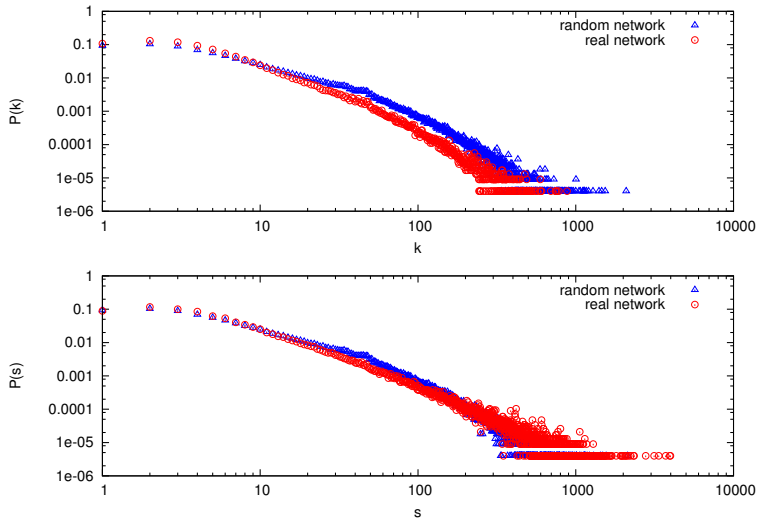

Figure 3. Real (red circles) and random (blue triangles) social network distributions of degrees (up) and strengths (bottom)

network, then built from those a random bipartite network following (Guillaume *et al.*, 2003), and then obtained our epistemic network by considering the concepts network projection. In this way we have a guarantee that our random model is more realistic than a simple random network with the same degree distribution as the real epistemic network. At this point we can also observe that the heterogeneous distribution of the degrees and the strengths is due to the heterogeneity of the degree distribution of the concepts in the original bipartite network, as observed for all bipartite networks in (Guillaume *et al.*, 2003). As Figure 2 shows, the so obtained random network has a similar but not identical degree distribution, and an almost identical strength distribution. The random network turns out to be in fact about three time denser than the real network (by construction only the sum of the weights needs to be the same in the two networks). This means that in our epistemic network there is a stronger tendency to repeat existing edges with respect to the null model, meaning that two concepts that have been used together in an article are likely to be used together again. This phenomenon accounts for the fact that once scientists find a relation between two concepts, there is then a tendency from the scientific communities to continue exploring and further analyzing it. The higher density has as a direct consequence that the unweighted average shortest path length is lower in the random network (2.37 vs 2.73) and the average clustering coefficient is higher (0.60 vs 0.46). Moreover we found that the real epistemic network has an optimal modularity value of 0.56, which is an order of magnitude higher than the corresponding value for the random network (0.03). These are the modularity values that correspond to the partition of the network that maximizes such measure, which was found using the Louvain algorithm (Blondel *et al.*, 2008). This interesting result indicates that the structure of the real epistemic

network is more structured than a random model of the same network could predict, meaning that over time concepts tend to form highly connected groups that we could identify as the sub-fields of the considered discipline, as observed also in (Herrera *et al.*, 2010).

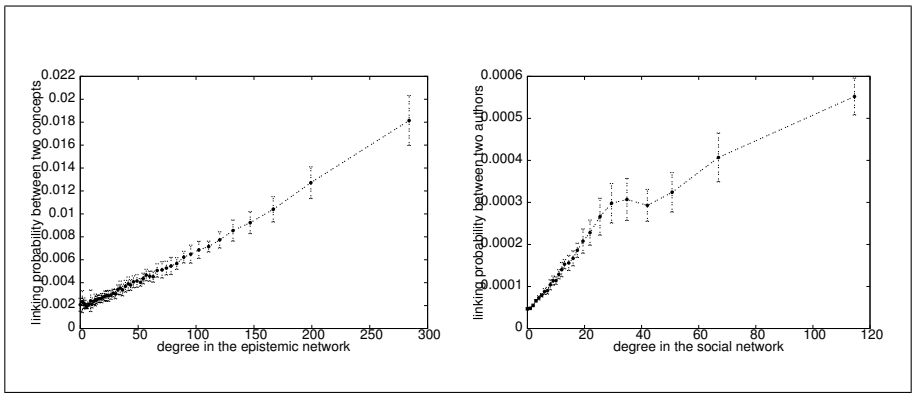

Figure 4. (Left) Linking probability between two concepts in the epistemic network in function of their degree. (Right) Linking probability between two authors in the social network in function of their degree. Values are averaged over all the considered years, and the error bars represent the 95% *confidence level interval.*

The epistemic network is therefore characterized by a heterogeneous degree distribution and a community structure. Since we have a longitudinal dataset we can try to understand what could be the microscopic dynamical mechanisms that give rise to these macroscopic features. The intuition is that what is at stake are a preferential attachment mechanism that would produce the scale free degree distribution, combined with a tendency to create short range links, which give raise to the community structure. Therefore we computed the probability that a new link would occur between two concepts in function of their degree in the epistemic network, and in function of their proximity in the network. The idea of proximity of two nodes in a network can be captured by different measures, such as the topological distance between them, the number of common neighbors, or some more refined index (for a complete review of such measures see (Lu *et al.*, 2011)). We decided to use the Jaccard index, i.e. the number of common neighbors over the cardinality of the union of the two set of neighbors (Jaccard, 1901), for its simplicity and robustness. Two nodes could in fact have a very short topological distance because there is just one other node connecting them even if they belong to different part of the network which are not very well connected, whereas the fraction of common neighbors gives us a better idea of how inter-connected the two nodes are. The link prediction problem can be stated as follows: given a snapshot of a network at time t, we would like to predict the links that will be added to the network from time t to a given future time t'. For further details see for example (Liben-Nowell *et al.*, 2007) and (Pujari *et al.*, 2012). Therefore we computed the probability of a new edge between two concepts that were never directly connected before, in function of the Jaccard similarity of their neighbours. In order to

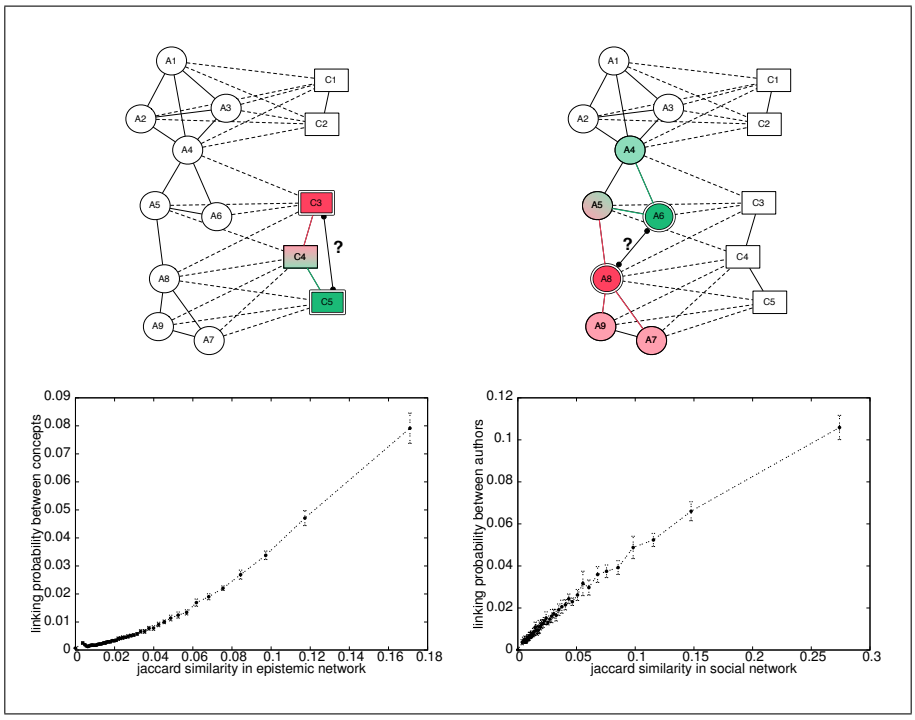

Figure 5. (Left) Linking probability between two concepts in the epistemic network in function of the Jaccard similarity of their neighbours. (Right) Linking probability between two authors in the social network in function of the Jaccard similarity of their neighbours. Values are averaged over all the considered years, and the error bars represent the 95% *confidence level interval.*

avoid artificially biased results, we restrict this analysis to the giant connected component of the network, which on average (over the years) covers more than 90% of the nodes. Also, when considering the network snapshots over time we introduced a life time for the links, since it would be unrealistic to consider a link still in place if two concepts (or two authors for the following part) haven't co-occurred any more for several years. So we consider a link between two nodes only if the last interaction they had is not older than 5 years, which seems like a realistic threshold since we verified in the data that 95% of concepts and authors have a time intervals between two successive interactions of under five years, and that 90% of them interact for less than five years overall. Figure 4 (left) shows indeed that the higher the degree in year y, the higher the probability of a new link in year $y + 1$, and Figure 5 (left) shows that the higher the similarity between two concepts in year y, the higher the probability that two previously unconnected concepts will be connected in year $y + 1$. The first result may account for the heterogeneity in the degree distribution, and the second one for the modular structure of the network.

For the social network we found similar results: a scale-free degree distribution (see Figure 3), an edge density about twice higher in the random network than in the real social network, and a modularity value of the real network higher than the one of the random one, even though the difference is not as marked as in the epistemic network but still significant (0.80 vs 0.38, i.e. about the double). Also in this case there is a growth in the probability of a new link between two authors that never collaborated before in function of their neighbors similarity, as shown in Figure 5 (right), indicating a tendency to collaborate with people our collaborators already worked with, as recently showed also in (Martin *et al.*, 2013).

Our main goal is to understand if the two networks under study show any sign of co-evolution. In order to do that we studied the probability of formation of a new edge in the social (resp. epistemic) network in function of the number of concepts (resp. authors) the two authors (resp. concepts) have in common (normalized again by the cardinality of the union of the two sets). Figure 6 shows that these probabilities grow as this measure of inter-network similarity grows, indicating that there is indeed a mutual co-evolution of the two networks. In particular if we look at the values we can observe that links in the epistemic network are predicted with the same probability by looking at the endogenous and exogenous similarity index, indicating that both play an important role in the evolution of the network. The values are different for what concerns the authors though, suggesting that the the social dimension plays the bigger role in the evolution of this socio-epistemic system.

These results show that the social and the epistemic network mutually influence each other, meaning that researchers choose their collaborators and the concepts they will work on in function of social factors (i.e. the proximity in the collaboration network), but also in function of epistemic factors such as the collaborator epistemic similarity. By performing further analysis in this direction we would like to be able to answer the question to which extent there is a tendency of two researchers collaborating because they know each other because they have common collaborators, and to which extent rather because they are experts of the same concepts.

5. Conclusions and Future Works

In this work we showed that the concept network of scientific fields such as physics has a non trivial structured characterized by a scale-free degree distribution and at the same time a tendency to local linking which tends to create a community structure that we can interpret as the structure in different sub-fields, characteristic of all scientific domains. Moreover we showed that there are clear indicators of the fact that the social and the epistemic networks mutually influence each other over time, giving rise to a complex socio-epistemic dynamics. We will further investigate these phenomena by building a model of network co-evolution in order to try to quantify the different contributions of the two dimensions, social and epistemic, in the evolution of scientific fields. The idea of the model is the following. For every article whose creation we want to simulate we pick a main author and then we implement a random walk in

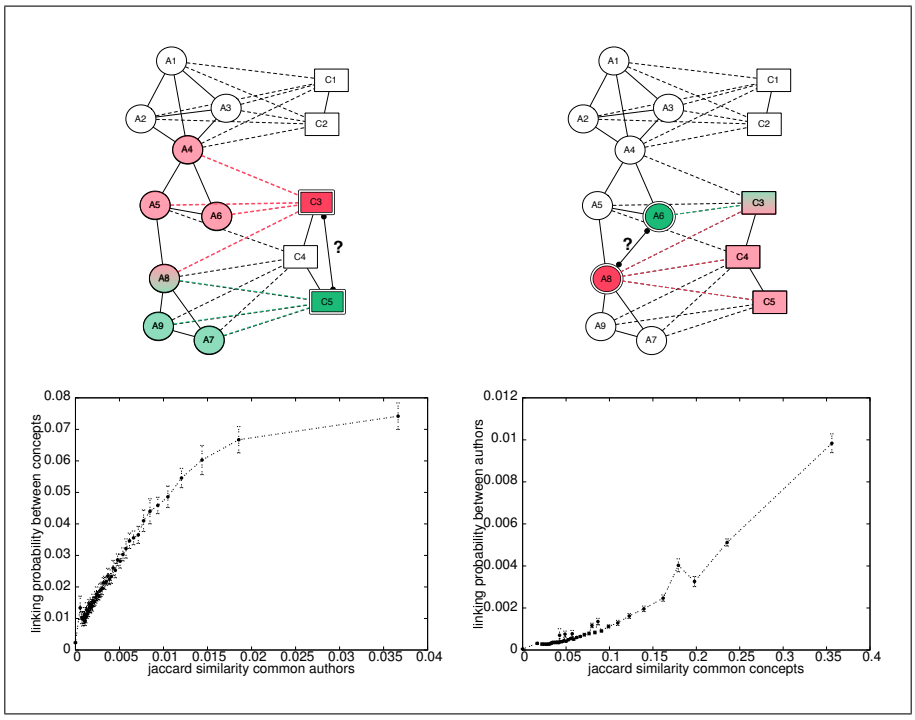

Figure 6. (Left) Linking probability between two concepts in the epistemic network in function of common authors Jaccard similarity. (Right) Linking probability between two authors in the social network in function of common concepts Jaccard similarity. Values are averaged over all the considered years, and the error bars represent the 95% *confidence level interval.*

the complete socio-epistemic network starting from her/him and then picking all the other authors and concepts that we cross during the walk until we reach the desired number of authors and concepts that we have drawn from the corresponding empirical distribution. The possible success of a model of this kind is indeed suggested by the results shown in Figure 7 suggesting that path lengths in the socio-espitemic network play a crucial role in determining the probability that an author use a given concept.

References

Newman, M. E. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Physical review E*, 64(1), 016131.

Newman, M. E. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical review E*, 64(1), 016132.

Herrera, M., Roberts, D. C., & Gulbahce, N. (2010). Mapping the evolution of scientific fields. PloS one, 5(5), e10355.

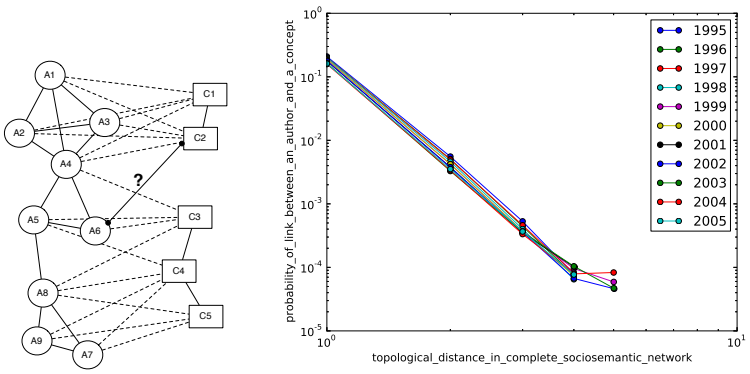

Figure 7. (Left) Linking probability that an author (here A6) will use a given concept (here C2) in function of topological distance computed in the common authors Jaccard similarity. The probability for an author to use a concept is exponentially decreasing with their topological distance in the hybrid socio-epistemic network.

Roth, C. (2005). Co-evolution in epistemic networks–Reconstructing social complex systems. *Structure and Dynamics*, 1(3).

Sun, X., Kaur, J., Milojevic, S., Flammini, A., and Menczer, F. (2013). Social Dynamics of Science. *Scientific reports*, 3.

Martin, T., Ball, B., Karrer, B., and Newman, M. E. J. (2013). Coauthorship and citation in scientific publishing. *arXiv preprint arXiv:1304.0473*.

Latapy, M., Magnien, C., and Vecchio, N. D. (2008). Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1), 31-48.

Guillaume, J. L., and Latapy, M. (2003). A realistic model for complex networks. *arXiv preprint cond-mat/0307095*.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008.

Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications, 390(6), 1150-1170.

P. Jaccard, Etude comparative de la distribution florale dans une portion des Alpes et des Jura, Bull. Soc. Vaud. Sci. Nat. 37 (1901) 547.

Leicht, E. A., Holme, P., & Newman, M. E. J. (2006). Vertex similarity in networks. Physical Review E, 73(2), 026120.

D. Liben-Nowell et al., The link-prediction problem for social networks JASIST, 2007, 58, 1019-1031.

M. Pujari et al., Link Prediction in Complex Networks by Supervised Rank Aggregation. ICTAI'12 : 24th International Conference on Tools with Artificial Intelligence, 2012