Expanding the OntoDM ontology with network analysis tasks and algorithms

Jan Kralj[1,2]
jan.kralj@ijs.si

Panče Panov[1]
pance.panov@ijs.si

Sašo Džeroski[1,2]
saso.dzeroski@ijs.si

[1]Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
[2]Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

ABSTRACT

This paper presents the first steps towards integrating concepts from the field of network analysis into the OntoDM ontology of data mining concepts. We have performed an extensive analysis of different subfieds of network analysis to provide a broad overview of the variety of tasks and algorithms that are encountered in the field. The main part of this work was to categorize the tasks and algorithms into a hierarchy that is consistent with the structure of OntoDM and which can systematically cover as many aspects of network analysis as possible. This work is a first step in the direction of OntoDM becoming an ontology that systematically describes not only data mining, but also network analysis. We believe that this work will encourage other researchers working in the filed to provide additional insight and further improve the integration of this field into OntoDM.

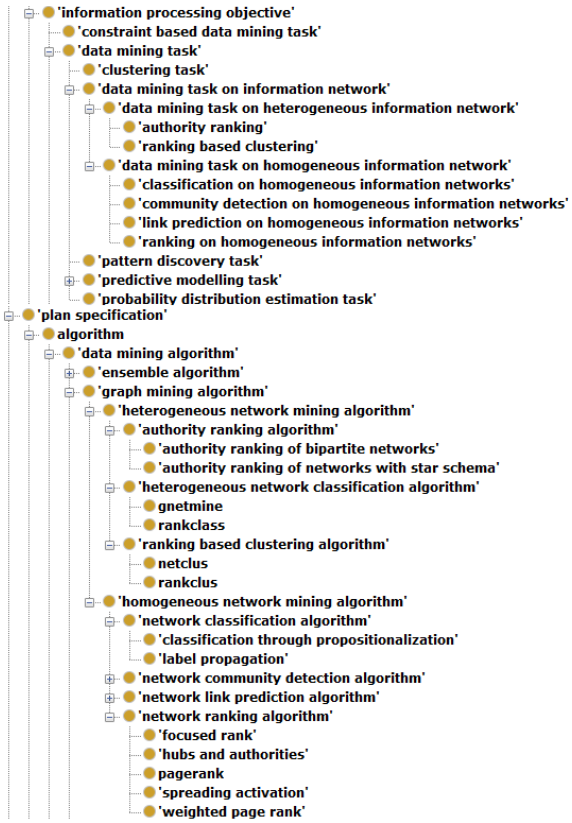
1. INTRODUCTION

Network analysis, is a large and quickly growing scientific discipline connected to physics, mathematics, social sciences and data mining. The tasks, tackled by the experts in the field, range from detecting communities in a given network, through predicting links in incomplete or time evolving networks, to ranking or classifying vertices of a given network. Most such tasks are analyzed in the context of information networks in which all nodes are treated equally, but in recent years, the concept of *heterogeneous* information networks [43], a generalization of standard information networks (which are then referred to as *homogeneous*), is gaining popularity.

OntoDM [35, 36] is a reference modular ontology for the domain of data mining. It is directly motivated by the need for formalizing the data mining domain and is designed and implemented by following ontology best practices and design principles. It includes the terms neccessary to describe different types of data, data mining tasks and approaches to solving these tasks. Among the key OntoDM classes are the classes representing datasets, data mining tasks, generalizations and algorithms themselves. The latter three classes are interconnected, as each data mining algorithm solves some data mining task by producing an output which is some type of generalization. In our work, we have expanded these classes to include tasks and algorithms that are found in the study of network analysis.

2. ONTOLOGY EXTENSION

This section presents an overview of the ontology classes that we have added to the OntoDM ontology. Shown in Figure 1, they consist mainly of subclasses of the classes *data mining task* and *data mining algorithm*.



```
  'information processing objective'
    'constraint based data mining task'
    'data mining task'
      'clustering task'
      'data mining task on information network'
        'data mining task on heterogeneous information network'
          'authority ranking'
          'ranking based clustering'
        'data mining task on homogeneous information network'
          'classification on homogeneous information networks'
          'community detection on homogeneous information networks'
          'link prediction on homogeneous information networks'
          'ranking on homogeneous information networks'
      'pattern discovery task'
      'predictive modelling task'
      'probability distribution estimation task'
'plan specification'
  algorithm
    'data mining algorithm'
      'ensemble algorithm'
      'graph mining algorithm'
        'heterogeneous network mining algorithm'
          'authority ranking algorithm'
            'authority ranking of bipartite networks'
            'authority ranking of networks with star schema'
          'heterogeneous network classification algorithm'
            gnetmine
            rankclass
          'ranking based clustering algorithm'
            netclus
            rankclus
        'homogeneous network mining algorithm'
          'network classification algorithm'
            'classification through propositionalization'
            'label propagation'
          'network community detection algorithm'
          'network link prediction algorithm'
          'network ranking algorithm'
            'focused rank'
            'hubs and authorities'
            pagerank
            'spreading activation'
            'weighted page rank'
```

Figure 1: The hierarchical structure of the main network analysis additions to OntoDM

As a subclass of the data mining task class, we have added a new class of data mining tasks, *data mining task on information network*, which includes all tasks encountered in our overview of the field. The tasks are first split into tasks that can be defined on a general (homogeneous or heterogeneous) information network and those that can only be defined on a heterogeneous network. Tasks on general networks, which constitute the majority of the new entries, include link prediction, community detection, ranking, and classification.

The classes, added under the *data mining algorithm* class, are gathered into a separate parent class *network analysis algorithm.* Each algorithm constitutes a leaf node in the hierarchy rooted in this parent class. The hierarchical structure of network analysis algorithms follows the hierarchy of tasks, described in the previous paragraph. Each presented algorithm solves a particular task which lies in a analogous part of the ontology – heterogeneous network analysis [43], link prediction algorithms [29, 2], community detection methods [12, 37], network ranking algorithms [11] and network classification algorithms. Furthermore, for each network analysis algorithm, we added a short description presenting the key concepts of the algorithm, as well as references to the paper in which the algorithm was presented. The descriptions and references are added as annotations to the classes in the ontology. Note that some terminal nodes of the hierarchy are actually instances of the classes, while other are proper classes that still need to be populated.

The final concept we added to the ontology was the concept of generalization specifications. Generalization specifications describe the types of output given by network analysis algorithms. Because the outputs of network analysis tasks are fundamentally different from outputs of traditional data mining algorithms, we decided to construct a hierarchy of generalization specifications, following the hierarchy of network analysis tasks and algorithms.

3. DATA MINING TASKS

In this section, we present the classes, added to the *data mining task* class. We present a short description of each class of data mining tasks that was added to OntoDM.

3.1 Data mining tasks on general networks

Data mining tasks on general networks are data mining tasks that can be formulated on any (homogeneous or heterogeneous) network. Commonly, different algorithms are used to perform the same task on homogeneous and heterogeneous networks.

Classification. Classification of network data is a natural generalization of classification tasks encountered in a typical machine learning setting. The problem formulation is simple: given a network and class labels for some of the vertices in the network, predict the class labels for the rest of the vertices in the network. The output of a classification task on a network is a function that predicts the class label of each vertex in the network.

Link prediction. While classification tasks try to discover new knowledge about network entities, link prediction focuses on unknown connections between the entities. The assumption is that not all network edges are known. The task of link prediction is to predict new edges that are missing or likely to appear in the future. The output of an algorithm solving a link prediction task is a function which provides a proximity measure for each pair of vertices in a network. The pairs with the highest proximity measure are then assumed to be the most likely candidates for predicted links.

Community detection. While there is a general consensus on what a network community is, there is no strict definition of the term. The idea is well summarized in the definition by Yang et al. [49]: a community is a group of network nodes, with dense links within the group and sparse links between groups. The output of a community detection algorithm is similar to the output of a clustering algorithm: a function that assigns each vertex in the network to a community.

Ranking. The objective of ranking in information networks is to assess the relevance of a given object either with regard to the whole graph or relative to some subset of vertices in the graph. In either case, the output of a network ranking algorithm is a function that assigns a score to each vertex of the network. The vertices with the highest score are then ranked the highest.

3.2 Data mining tasks on heterogeneous networks

Data mining tasks on heterogeneous networks are tasks that can only be formulated on a heterogeneous network. Unlike the tasks on general networks, these tasks have only been addressed in recent years.

Authority ranking. Sun and Han [43] introduce *authority ranking* to rank the vertices of a heterogeneous network with either a bipartite structure or a star network schema in which one vertex type is central in that all network edges start or end on a vertex of this central type. The task of authority ranking is to rank vertices in each (not necessarily all) vertex type separately, and the output of the task is a collection of functions, each assigning a score to only vertices of a certain type.

Ranking based clustering. While both ranking and clustering can be performed on heterogeneous information networks, applying only one of the two may sometimes lead to results which are not truly informative. For example, simply ranking authors in a bibliographic network may lead to a comparison of scientists in completely different fields of work which may not be comparable. Sun and Han [43] propose joining the two seemingly orthogonal approaches to information network analysis (ranking and clustering) into one, in which vertices are simultaneously assigned to a cluster and given a score to rank them within the cluster.

4. ALGORITHMS

This section presents the algorithms that are classified in the modified ontology. The classification hierarchy of network analysis algorithms is similar to the hierarchy of data mining tasks, described in Section 3.

4.1 General network mining algorithms

This section describes the algorithms that can be used to solve data mining tasks on general networks.

Community detection algorithms. The classification of community detection algorithms on networks follows the classification of algorithms, described in the surveys by Fortunato [12] and Plantié and Crampes [37]. The algorithms can be split into several classes based on the underlying idea that guides the algorithms. It must be noted that a strict split of the different methods is impossible as different methods are not developed in isolation. For example, many methods that are not strictly classified as modularity based algorithms still use the concept of modularity in one of their steps.

Divisive algorithms. Divisive algorithms are algorithms that find a community structure of a network by iteratively removing edges from the network. As edges are removed, the network decomposes into disconnected components. The decomposition pattern forms a hierarchical clustering over the set of all vertices in the network. The most widely used such algorithm is the Girvan Newman algorithm [16], which removes the edges in the network with the largest centrality measure, arguing edges which are more central to a graph are the edges that cross communities. An alternative algorithm is the Radicchi algorithm which calculates the edge clustering coefficient of edges to calculate which edges must be removed. Here, the intuition is that edges between communities belong to fewer cycles than edges within communities.

Modularity based algorithms. Modularity based algorithms form the majority of community detection algorithms. While the concept of modularity (first defined in Newman and Girvan [33]) is used in almost all algorithms at some point (especially to determine the best clustering from a hierarchical clustering of nodes), the algorithms in this class use modularity more centrally than other algorithms. The most prominent such methods are the Louvain algorithm [5] and the Newman greedy algorithm [33]. Other methods include variations of the greedy algorithm [46], using simulated annealing [18], spectral optimization of modularity via a modularity matrix [32, 31] or via the graph adjacency matrix [47], and deterministic optimization approaches [10].

Spectral algorithms. Spectral algorithms find communities in network by analyzing eigenvectors of matrices, derived from the network. The community structure is extracted either from the eigenvectors of the Laplacian matrix of the network [9] or from the stochastic matrix of the network [6]. In both cases, algorithms assume that eigenvectors, extracted from the network, will have similar values on indices that belong to network vertices in the same community. The computation of several eigenvectors belonging to the largest eigenvalues is first performed. The eigenvectors form a set of coordinates of points, each belonging to one network vertex. Clustering of the points then corresponds to community detection of network vertices.

Random walk based algorithms. Random walk based algorithms are algorithms that use the concept of a random walker on a network to perform community detection. The methods use a random walker model to determine the similarities of network vertices and then use either a divisive [51] or an agglomerative [52, 38] approach to construct a hierarchical clustering of the nodes.

Link prediction algorithms. Link prediction algorithms are presented in survey papers by Lü and Zhou [29] and Al Hasan and Zaki [2]. These surveys present a similar hierarchy of link prediction algorithms, which we also used in the construction of the ontology. All presented algorithms calculate a proximity measure between two vertices. They do so in 3 distinct ways, described below.

Similarity based algorithms. Similarity based algorithms calculate the proximity of two vertices in the network either from their neighborhoods (*local similarity based algorithms*) or from the way the two vertices fit into the overall network structure (*global similarity based algorithms*). Local similarity based algorithms are further divided into common neighbor based algorithms and vertex degree based algorithms. The first class of algorithms computes the similarity between two vertices purely from the number of neighbors of each node, and the number of common neighbors, while the second class also takes the degrees of both nodes into account. The most widely used algorithm in this class is the algorithm for calculating the Adamic-Adar proximity measure [1]. Other proximity measures listed are the common neighbors [30], the hub depressed and hub promoted indices [39], the Jaccard index, the Leicht-Holme-Newman index [28], the Salton index [40], the Sorensen index [42] and the preferential attachment index [3]. Unlike local similarity based algorithms, global similarity based algorithms use the entire network structure to calculate the proximity between two network vertices. The algorithms include the Katz index [25], the random walk with restart [41], the SimRank [22], the average commute time index [26] and the matrix forest index [7].

Probability based algorithms. Probabilistic algorithms for link prediction use various techniques to estimate the probability that a pair of vertices should be connected. These maximum likelihood methods, like the hierarchical structure model [8] and the stochastic block model [19], and probabilistic models, like the probabilistic relational model [15], probabilistic entity relationship model [20] and stochastic relational models [50].

Network ranking algorithms. The classification of network ranking algorithms adopted in this work was guided by the paper by Duhan et al. [11]. However, this paper is not as detailed as the survey papers for the link prediction and community detection tasks. The paper focuses on the classification of web pages and describes several algorithms for ranking vertices in a network. For this work, only the methods that deal with ranking nodes in a network were used. The methods include the famous PageRank algorithm [34] used by the Google search engine and a weighted version of the PageRank method called the Weighted PageRank [48], as well as the related Hubs and Authorities method [27]. Another method to rank nodes in the network is to use centrality measures. To construct a collection of network centrality measures, we followed the lecture given by dr. Cecilia Mascolo [1]. The network centrality measures listed in the ontology are Freeman's Network Centrality [14], betweenness centrality [13], closeness centrality [4] and the Katz centrality measure [25].

Network classification algorithms. The most widely used network classification algorithm is the label propagation algorithm [52]. Another algorithm based on nwtwork propositionalization [17] can also be used to classify nodes in a homogeneous network.

[1] https://www.cl.cam.ac.uk/teaching/1314/L109/stna-lecture3.pdf

4.2 Heterogeneous network mining algorithms

This section describes the algorithms used to solve the data mining tasks, described in Section 3.2.

Authority ranking. Authority ranking, as presented in [43], can be adressed by the algorithms for authority ranking in networks with a bipartite structure and in networks with a network schema.

Ranking based clustering. Ranking based clustering, as presented in [43], is adressed similarly to authority ranking. Sun et al. [44] present the algorithm RankClus, which performs ranking based clustering on bipartite networks, and Sun et al. [45] present the NetClus algorithm which tackles the same task on networks with a star network schema.

Classification in heterogeneous networks. The algorithms in this class can be used to classify nodes in a heterogeneous network. They are the algorithm by Grčar et al. [17] which uses network propositionalization to classify network vertices, the RankClass [23], the GNetMine [24] algorithm and the heterogeneous network propagation algorithm [21].

5. CONCLUSION AND FURTHER WORK

The field of network analysis is a rich and complex field. This work presents a starting point for integrating the descriptions of tasks and algorithms ito OntoDM. In the future, we wish to add several subfields of network analysis that were not analyzed in this paper, such as the analysis of data enriched networks, time evolving networks and a separate analysis of community detection algorithms for directed networks. Furthermore, the ontology can be expanded to include example datasets on which algorithms can be tested, as well as evaluation metrics to examine the performance of various algorithms.

Acknowledgement. We would like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

References

[1] Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the Web. *Social Networks*, 25(3):211–230.

[2] Al Hasan, M. and Zaki, M. J. (2011). A survey of link prediction in social networks. In Aggarwal, C. C., editor, *Social Network Data Analytics*, pages 243–275. Springer.

[3] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

[4] Bavelas, A. (1950). Communication patterns in task-oriented groups. *Journal of the Acoustical Society of America*.

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

[6] Capocci, A., Servedio, V. D., Caldarelli, G., and Colaiori, F. (2005). Detecting communities in large networks. *Physica A: Statistical Mechanics and its Applications*, 352(2):669–676.

[7] Chebotarev, P. Y. and Shamis, E. (1997). A matrix-forest theorem and measuring relations in small social group. *Avtomatika i Telemekhanika*, (9):125–137.

[8] Clauset, A., Moore, C., and Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101.

[9] Donetti, L. and Munoz, M. A. (2004). Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(10):P10012.

[10] Duch, J. and Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Physical Review E*, 72(2):027104.

[11] Duhan, N., Sharma, A., and Bhatia, K. K. (2009). Page ranking algorithms: A survey. In *2009 IACC Advance Computing Conference*, pages 1530–1537. IEEE.

[12] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75–174.

[13] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

[14] Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239.

[15] Friedman, N., Getoor, L., Koller, D., and Pfeffer, A. (1999). Learning probabilistic relational models. In *16th International Joint Conference on Artificial Intelligence*, volume 99, pages 1300–1309.

[16] Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

[17] Grčar, M., Trdin, N., and Lavrač, N. (2013). A methodology for mining document-enriched heterogeneous information networks. *The Computer Journal*, 56(3):321–335.

[18] Guimera, R. and Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900.

[19] Guimerà, R. and Sales-Pardo, M. (2009). Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078.

[20] Heckerman, D., Meek, C., and Koller, D. (2007). Probabilistic entity-relationship models, PRMs, and plate models. In Getoor, L. and Taskar, B., editors, *Introduction to Statistical Relational Learning*, pages 201–238. MIT Press.

[21] Hwang, T. and Kuang, R. (2010). A heterogeneous label propagation algorithm for disease gene discovery. In *10th SIAM International Conference on Data Mining*, pages 583–594.

[22] Jeh, G. and Widom, J. (2002). SimRank: a measure of structural-context similarity. In *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 538–543. ACM.

[23] Ji, M., Han, J., and Danilevsky, M. (2011). Ranking-based classification of heterogeneous information networks. In *17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1298–1306. ACM.

[24] Ji, M., Sun, Y., Danilevsky, M., Han, J., and Gao, J. (2010). Graph regularized transductive classification on heterogeneous information networks. In *2010 European Conference on Machine Learning and Knowledge Discovery in Databases*, volume 1, pages 570–586. Springer-Verlag.

[25] Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

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

[27] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

[28] Leicht, E., Holme, P., and Newman, M. E. (2006). Vertex similarity in networks. *Physical Review E*, 73(2):026120.

[29] Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170.

[30] Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102.

[31] Newman, M. E. (2006a). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104.

[32] Newman, M. E. (2006b). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

[33] Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113.

[34] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

[35] Panov, P., Džeroski, S., and Soldatova, L. N. (2008). OntoDM: An ontology of data mining. In *2008 IEEE International Conference on Data Mining Workshops*, pages 752–760. IEEE Computer Society.

[36] Panov, P., Soldatova, L., and Džeroski, S. (2014). Ontology of core data mining entities. *Data Mining and Knowledge Discovery*, 28(5-6):1222–1265.

[37] Plantié, M. and Crampes, M. (2013). Survey on social community detection. In Ramzan, N. e. a., editor, *Social Media Retrieval*, pages 65–85. Springer.

[38] Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks. In *20th International Symposium on Computer and Information Sciences*, pages 284–293. Springer.

[39] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555.

[40] Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. Mcgraw-Hill College.

[41] Shang, M.-S., Lü, L., Zeng, W., Zhang, Y.-C., and Zhou, T. (2009). Relevance is more significant than correlation: Information filtering on sparse data. *Europhysics Letters*, 88(6):68008.

[42] Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter*, 5:1–34.

[43] Sun, Y. and Han, J. (2012). *Mining Heterogeneous Information Networks: Principles and Methodologies*. Morgan & Claypool Publishers.

[44] Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., and Wu, T. (2009a). RankClus: integrating clustering with ranking for heterogeneous information network analysis. In *2009 International Conference on Extending Database Technology*, pages 565–576.

[45] Sun, Y., Yu, Y., and Han, J. (2009b). Ranking-based clustering of heterogeneous information networks with star network schema. In *15th SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806.

[46] Wakita, K. and Tsurumi, T. (2007). Finding community structure in mega-scale social networks:[extended abstract]. In *16th international Conference on the World Wide Web*, pages 1275–1276. ACM.

[47] White, S. and Smyth, P. (2005). A spectral clustering approach to finding communities in graph. In *5th SIAM International Conference on Data Mining*, volume 5, pages 76–84. SIAM.

[48] Xing, W. and Ghorbani, A. (2004). Weighted pagerank algorithm. In *2nd Annual Conference on Communication Networks and Services Research*, pages 305–314. IEEE.

[49] Yang, B., Liu, D., and Liu, J. (2010). Discovering communities from social networks: Methodologies and applications. In Furht, B., editor, *Handbook of Social Network Technologies and Applications*, pages 331–346. Springer.

[50] Yu, K., Chu, W., Yu, S., Tresp, V., and Xu, Z. (2006). Stochastic relational models for discriminative link prediction. In *2006 Conference on Advances in Neural Information Processing Systems*, pages 1553–1560.

[51] Zhou, H. (2003). Network landscape from a Brownian particle's perspective. *Physical Review E*, 67(4):041908.

[52] Zhou, H. and Lipowsky, R. (2004). Network Brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities. In *2004 International Conference on Computational Science*, pages 1062–1069. Springer.