Mesures d'influence via les indicateurs de centralité dans les réseaux sociaux

MARAMI 2015, 14-16 octobre Nîmes

Mesures d'influence via les indicateurs de centralité dans les réseaux sociauxs

Oualid Benyahia
Christine Largeron

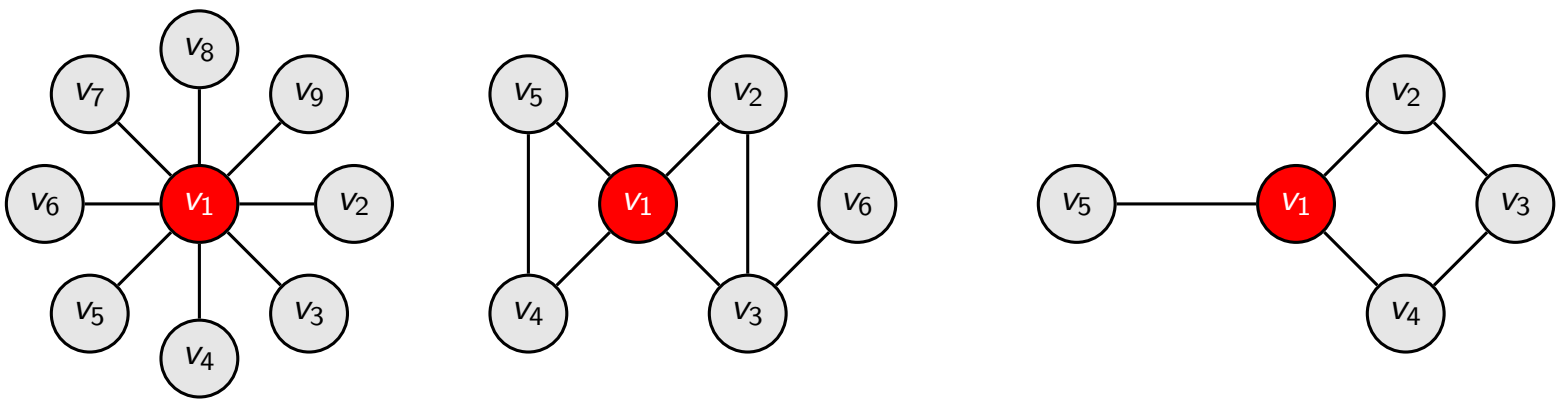
Laboratoire Hubert Curien, Université Jean-Monnet Saint-Etienne

MARAMI 16 Octobre 2015

Outline

Context

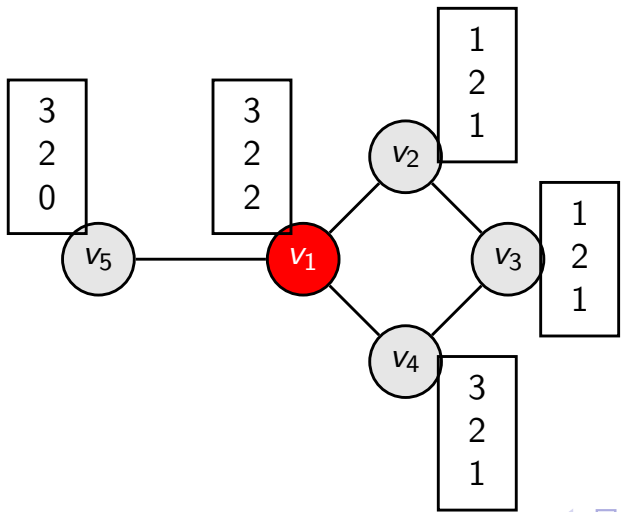
- Identify actors with important roles in a network.
  - Information recommendation
  - Viral marketing
- Several indicators initially introduced in social network analysis as measures of centrality : *Degree*, *Closeness*, *Betweenness*,$\cdots$ [Freeman, 1979, Wasserman and Faust, 1994]
- Centrality measures evaluate the importance of an actor considering only his structural position in the network.



Context

- Social networks become more complex and heterogeneous with information qualifying the users.
- Information network [Sun et al. 2009]: "A network where each node represents an entity (e.g., actor in a social network) and each link (e.g., tie) a relationship between entities
  - Each node/link may have attributes, labels, and weights
  - Link may carry rich semantic information"
- Information networks can be represented as attributed and weighted graphs.

$\Rightarrow$ In our work each node is associated with features that can be used to asses its importance or influence.



Objectif

Our aim :

- Adapt the classical centrality indicators to deal with complex graphs where:
  - Links weight quantify the intensity of the relations between nodes.
  - Nodes are characterized by attributes or features describing their notoriety.
- Integrate numerical attributes that are more likely to characterize the importance or the influence of an actor.
  - Age
  - Popularity
  - Experience
  - Frequency of activity
  - Degree of involvement in a community
  - ...

Attributed networks

Two definitions of attributed networks

1. **Definition 1** [Zhou et al., 2009] :

An attributed graph is a graph $G = (V, E)$ where each node in V is assigned a vector of attributes.

2. **Definition 2** [Yin et al., 2010, Gong et al., 2011]:

An attributed graph is defined by:

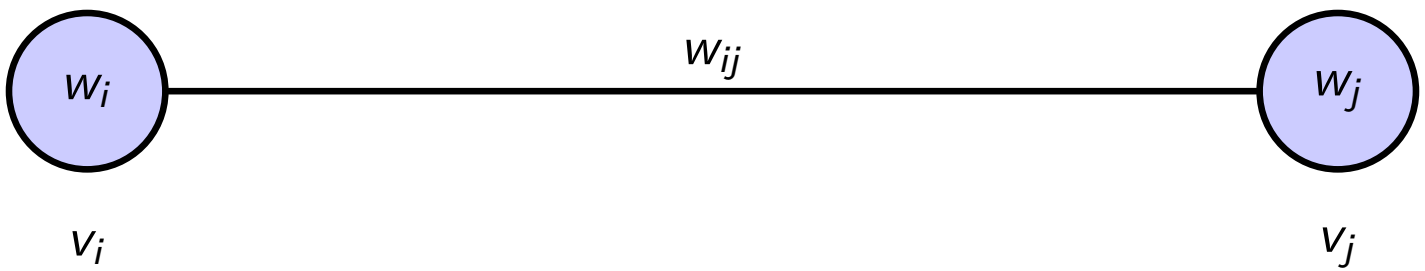
- a simple graph $G = (V, E)$ describing links between nodes.
- and a bipartite graph $G = (V \cup V_a, E_a)$ describing the relations E_a between nodes in V and their corresponding attributes in V_a.

Attributed networks

We retain the first definition [Zhou et al., 2009].

Definition

- Graph $G = (V, E)$ where V is the set of nodes and $E \subset V \times V$ is the set of edges.
- Each node $v_i \in V$ is characterized with a numerical vector $Y^i = (y_1^i, y_2^i, \ldots, y_L^i)^T$ describing its notoriety.
- The global attribute weight of a node v_i is computed by means of the norm of the vector Y^i: $w_i = \| Y^i \|$.
- The weight w_{ij}, of the edge between two nodes v_i and v_j, describes the intensity of their relation.

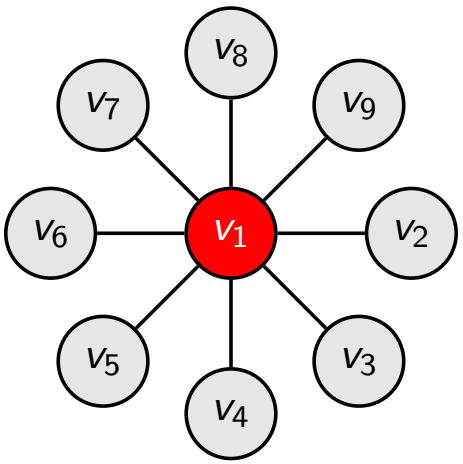


Degree measures

- Measures the relative importance of a node by counting its direct links in the graph, after normalization by the size of the network [Freeman, 1979].

Degree centrality

$$Degree(v_i) = \frac{deg(v_i)}{|V| - 1} \quad (1)$$



Degree measures adapted to weighted graphs

- Extended to the case of weighted and directed graphs [Barrat et al., 2004]

Weighted edge degree centrality

$$WEDegree(v_i) = \sum_{v_j \in out(v_i)} w_{ij} \quad (2)$$

- Get a balance between number of links and their weights [Opsahl et al., 2010]

Weighted edge Opshal degree centrality

$$WEOpsahlDegree = ((deg_{out}(v_i))^{(1-\alpha)} \cdot \left(\sum_{v_j \in out(v_i)} w_{ij} \right)^{\alpha}) \quad (3)$$

Degree measures adapted to attributed graphs

- For attributed graphs, the degree centrality is weighted by the general node notoriety given by its weight w_i.

$$WNDegree(v_i) = w_i \cdot Degree(v_i) \quad (4)$$

$$WNEDegree(v_i) = w_i \cdot WEDegree(v_i) \quad (5)$$

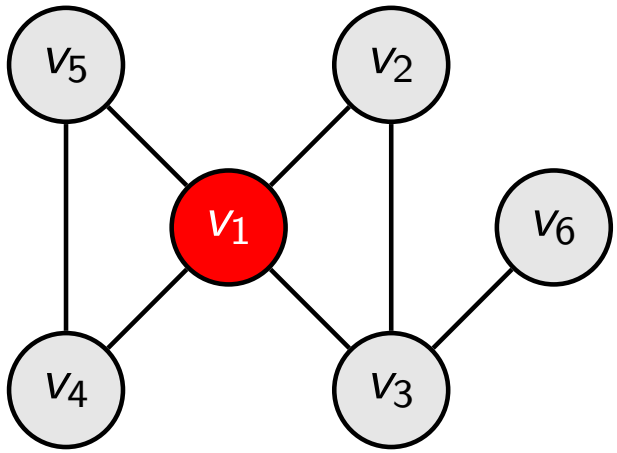
$$WNEOpsahlDegree(v_i) = w_i \cdot WEOpsahlDegree \quad (6)$$

Closeness measures

- For Closeness measures a node is considered as important if it can rapidly reach the other nodes of the graph [Wasserman and Faust, 1994, Hakimi, 1964].
- The usual measure is defined by the inverse of the sum of the geodesic distances of a given node to others nodes:

Closeness centrality

$$CCentr(v_i) = \frac{1}{\sum_{\substack{v_j \in V \\ j \neq i}} |ShortPath(v_i, v_j)|} \tag{7}$$



Closeness measures adapted to weighted graphs

- For weighted edge graphs, the closeness is computed by the sum of the weighted geodesic distances between a given node and the other nodes of the network.

Adapted weighted edge closeness centrality

$$CWECentr(v_i) = \sum_{\substack{v_j \in V \\ j \neq i}} \frac{\sum_{e \in ShortPath(v_i, v_j)} w(e)}{|ShortPath(v_i, v_j)|} \quad (8)$$

where $w(e)$ is the weight of the edge $e \in E$.

Closeness measures adapted to attributed graphs

- For attributed graphs, the closeness centrality is weighted by the general node notoriety given by its weight w_i.

Adapted closeness centrality for attributed graph

$$CWNCentr(v_i) = w_i \cdot CCentr(v_i) \tag{9}$$

- For attributed weighted graphs, the closeness centrality is weighted by the general node notoriety given by its weight.

Adapted weighted edge closeness centrality for attributed graph

$$CWNECentr(v_i) = w_i \cdot CWECentr(v_i) \tag{10}$$

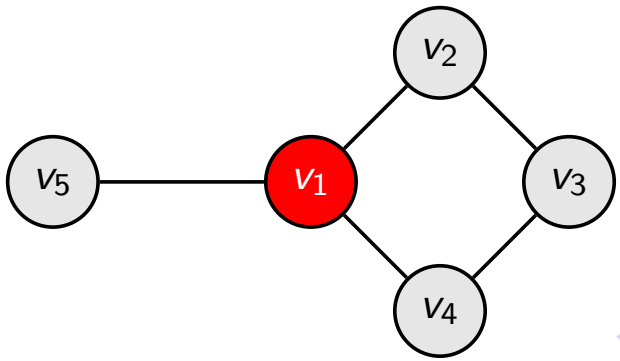
Betweenness measures

- In betweenness measure, a node is important, if it is located on a great number of paths between other nodes.
- Formally, it is equal to the normalized number of shortest paths between all pairs of nodes that pass through that node [Wasserman and Faust, 1994, Freeman, 1979]

Betweenness centrality

$$BCentr(v_i) = \sum_{\substack{(v_k, v_j) \in V \times V \\ i \neq k \neq j}} |g_{kj}(v_i)| / |g_{kj}| \qquad (11)$$

where $g_{kj}(v_i)$ denote the set of the shortest paths between nodes v_k and v_j that pass through v_i and g_{kj} the set of all shortest paths between the nodes v_k and v_j.



Betweenness measures adapted to weighted graphs

- For weighted edge graphs, we adapt the betweenness measure to consider weights of the shortest paths passing through the node.

Adapted weighted edge betweenness centrality

$$BWECentr(v_i) = \sum_{\substack{(v_k,v_j)\in V\times V \\ i\neq j\neq k}} \frac{\sum_{S\in g_{kj}(v_i)} \sum_{e\in S} w(e)}{\sum_{S\in g_{kj}} \sum_{e\in S} w(e)} \quad (12)$$

Betweenness measures adapted to attributed graphs

- When the nodes are described by a set of attributes, the measure is weighted by the node's weight w_i.

Adapted betweenness centrality for attributed graph

$$BWNCentr(v_i) = w_i \cdot BCentr(v_i) \tag{13}$$

- For attributed weighted graphs, the betweenness centrality is weighted by the general node notoriety given by its weight.

Adapted weighted edge betweenness centrality for attributed graph

$$BWNECentr(v_i) = w_i \cdot BWECentr(v_i) \tag{14}$$

Authority measures

- Eigenvector centrality is based on the idea that the score of a node is higher if it is connected to nodes having a high score than if it is connected to nodes with a low score
  [Bonacich and Paulette, 2001, Ghosh and Lerman, 2010].
- Usually, the *eigenvector centrality* is recursively computed by the formula :

Eigenvector centrality

$$EVCentr(v_i) = \frac{1}{\lambda_1} \cdot \sum_{v_j \in V} a_{ij} \cdot EVCentr(v_j) \tag{15}$$

where $A = \{a_{ij}\}$ is the adjacency matrix of the graph and λ_1 is the largest eigenvalue obtained as solution of the equation $AX = \lambda X$.

- For weighted edge graphs, the measure *EVWNECentr* is computed on the weighted adjacency matrix of the weighted graph.

Authority measures

- The *PageRank* is a variant of *eigenvector centrality* [Brin and Page, 1998, Ghosh and Lerman, 2010].
- It was initially introduced to measure the popularity of Web pages and is usually defined by the equation:

PageRank centrality

$$PRankCentr(v_i) = (1 - \beta) \cdot W_0 + \beta \cdot \sum_{v_j \in V} a_{ji} \cdot \frac{PRankCentr(v_j)}{deg(v_j)} \quad (16)$$

where W_0 is generally fixed uniformly and is equal to $\frac{1}{|V|}$ for all nodes and β is a damping factor.

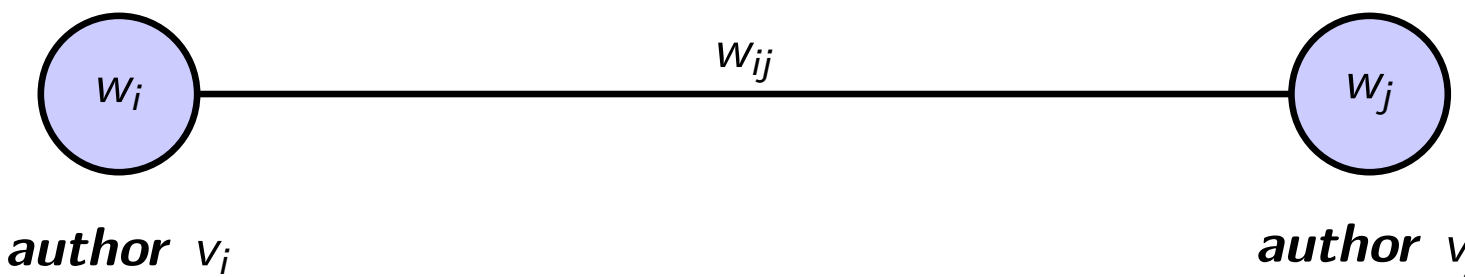
- In the case of attributed graph, we adopt a custom formulation *PRankWNECentr* of the *PageRank* in which the nodes weights w_i computed on the attributes are used instead of the uniform weights W_0 [Sergey et al., 1998, Jeh and Widom, 2003].

Datasets

- Academic coauthor networks [Tang et al., 2009] extracted from the *Arnetminer* academic research system.
- Dataset concerns 640134 authors and 1554643 coauthor relation on different topics.
- Topic distributions of authors and papers are discovered using a statistical topic modeling approach, Author-Conference-Topic (ACT) model [Tang et al., 2008].
- The ACT approach automatically extracts topics and assigns a topic distribution to each author and to each paper.
- Each topic graph describe co-publications in one topic.

Datasets

- Three attributed graphs were used, related respectively to the three topics : *Data mining*, *Information Retrieval (IR)* and *Bayesian Network*.
- An undirected graph $G = (V, E)$ is constructed with nodes set V representing authors and edge set E representing the co-publications, related to a topic, between different authors.
  - Each author $v_i \in V$ is characterized by an attribute w_i corresponding to the number of his publications.
  - The set E of edges represents the co-publication links weighted by the number of articles w_{ij} co-written by two authors v_i and v_j.



Datasets

Table: Statistics on the co-publications graphs.

Graphs	number of Nodes	number of edges
Data-Mining	679	1687
Information Retrieaval (IR)	657	1907
Bayesian Network	554	1238

Evaluation

- We consider two indicators as ground truth references of the author's influence extracted from Arnetminer [1] :
  - The *H-index* of an author [Hirsch, 2005].
  - The number of *Citations* of author publications.
- The ranking of the top 20 most important authors according to a given measure is compared to the ranking of the top most important authors provided by the two ground truth indicators: the **H-index** and the number of **Citations**.

[1] http://arnetminer.org/person-ranklist/hindex/89.

Evaluation

- The accuracy is measured by mean of the **Jaccard index** and the **Precision** and **Recall** :.

Jaccard index

$$Jaccard(L_{20}^m, L_{20}^*) = \frac{|L_{20}^m \cap L_{20}^*|}{|L_{20}^m \cup L_{20}^*|} \quad (17)$$

$$Precision(L_{20}^m, L_{20}^*) = \frac{|L_{20}^m \cap L_{20}^*|}{|L_{20}^m|} \quad (18)$$

$$Recall(L_{20}^m, L_{20}^*) = \frac{|L_{20}^m \cap L_{20}^*|}{|L_{20}^*|} \quad (19)$$

where L_{20}^m list of the top 20 best ranked authors according to a centrality measure m, L_{20}^* the list of the top 20 best ranked authors given by the reference indicator **H-index** (respectively number of **Citations**).

Note that $|L_{20}^m| = |L_{20}^*| \Rightarrow Precision(L_{20}^m, L_{20}^*) = Recall(L_{20}^m, L_{20}^*)$

Results

Table: Score results of influence measures on Data-mining graph.

	J(H-Index)	J(Citations)	P/R(H-Index)	P/R(Citations)
Degree	0.3333	0.1765	0.5	0.3
WNDegree(*)	0.3333	0.1765	0.5	0.3
WEDegree	0.3333	**0.2121**	0.5	**0.35**
WEOpsahlDegree	0.3333	**0.2121**	0.5	**0.35**
WNEDegree(*)	0.3793	**0.2121**	0.55	**0.35**
WNEOpsahlDegree(*)	0.3793	**0.2121**	0.55	**0.35**
CCentr	0.3333	0.1765	0.5	0.3
CWNCentr(*)	0.3333	**0.2121**	0.5	**0.35**
CWECentr(*)	0.0	0.0256	0.0	0.05
CWNECentr(*)	0.25	0.1765	0.4	0.3
BCentr	0.3333	0.1429	0.5	0.25
BWNCentr(*)	0.3793	0.1765	0.55	0.3
BWECentr(*)	0.2903	0.1765	0.45	0.3
BWNECentr(*)	0.3793	0.1765	0.55	0.3
PRankCentr	**0.4286**	0.1765	**0.6**	0.3
PRankWNECentr(*)	0.2903	0.1765	0.45	0.3
EVCentr	0.2121	0.1111	0.35	0.2
EVWNECentr(*)	**0.4286**	**0.2121**	**0.6**	**0.35**

- Proposed measures are highlighted.
- Measures which obtained the best accuracy are emphasized with bold font.

Results(2)

Table: Score results of influence measures on Information Retrieval (IR) graph.

	J(H-Index)	J(Citations)	P/R(H-Index)	P/R(Citations)
Degree	0.25	0.2121	0.4	0.35
WNDegree(*)	**0.4286**	**0.3333**	**0.6**	**0.5**
WEDegree	0.25	0.2121	0.4	0.35
WEOpsahlDegree	0.2903	0.25	0.45	0.4
WNEDegree(*)	0.2903	0.2121	0.45	0.35
WNEOpsahlDegree(*)	0.3793	0.2903	0.55	0.45
CCentr	0.25	0.2121	0.4	0.35
CWNCentr(*)	0.3793	0.2903	0.55	0.45
CWECentr(*)	0.0	0.0	0.0	0.0
CWNECentr(*)	0.2121	0.1765	0.35	0.3
BCentr	0.1765	0.1429	0.3	0.25
BWNCentr(*)	0.3793	0.2903	0.55	0.45
BWECentr(*)	0.25	0.2121	0.4	0.35
BWNECentr(*)	0.3793	0.2903	0.55	0.45
PRankCentr	0.25	0.2121	0.4	0.35
PRankWNECentr(*)	0.3793	0.25	0.55	0.4
EVCentr	0.25	0.2121	0.4	0.35
EVWNECentr(*)	0.2903	0.25	0.45	0.4

- Proposed measures are highlighted.
- Measures which obtained the best accuracy are emphasized with bold font.

Results(3)

Table: Score results of influence measures on Bayesian Networks graph.

	J(H-Index)	J(Citations)	P/R(H-Index)	P/R(Citations)
Degree	0.2121	0.1429	0.35	0.25
WNDegree(*)	0.3333	0.25	0.5	0.4
WEDegree	0.25	0.1765	0.4	0.3
WEOpsahlDegree	0.2903	0.2121	0.45	0.35
WNEDegree(*)	**0.3793**	0.2903	**0.55**	0.45
WNEOpsahlDegree(*)	**0.3793**	0.2903	**0.55**	0.45
CCentr	0.2121	0.1765	0.35	0.3
CWNCentr(*)	0.3333	0.25	0.5	0.4
CWECentr(*)	0.0526	0.0256	0.1	0.05
CWNECentr(*)	0.2903	0.2121	0.45	0.35
BCentr	0.2903	0.25	0.45	0.4
BWNCentr(*)	**0.3793**	**0.3333**	**0.55**	**0.5**
BWECentr(*)	0.25	0.25	0.4	0.4
BWNECentr(*)	0.3333	0.25	0.5	0.4
PRankCentr	0.3333	0.25	0.5	0.4
PRankWNECentr(*)	0.25	0.2121	0.4	0.35
EVCentr	0.1111	0.0811	0.2	0.15
EVWNECentr(*)	0.3333	0.25	0.5	0.4

- Proposed measures are highlighted.
- Measures which obtained the best accuracy are emphasized with bold font.

Conclusion

Our contribution :

- Variants of centrality measures suited to infer the important users in weighted graphs with node attributes.
- For Degree, betweenness, closeness and eigenvector, the results are improved when the weights of the links are considered.
  - the improvement is important if we take into account the attributes that describe the nodes, in addition of the weights of the links.

- Among our proposition, mainly :
  - The degree centrality variants **WNEDegree** and **WNEOpsahlDegree** that incorporate the links weights and nodes attributes, give a good estimation of the authors influence.
  - They are efficient in terms of processing times.

References I

Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2004).

The architecture of complex weighted networks.
Proceedings of the National Academy of Sciences of the United States of America, 101(11):3747–3752.

Bonacich, P. and Paulette, L. (2001).

Eigenvector-like measures of centrality for asymmetric relations.
Social Networks, 23(3):191–201.

Brin, S. and Page, L. (1998).

The anatomy of a large-scale hypertextual web search engine.
In *Proceedings of the seventh international conference on World Wide Web 7*, WWW7, pages 107–117.

Freeman, L. C. (1979).

Centrality in social networks: Conceptual clarification.
Social Networks, 1(3):215–239.

Ghosh, R. and Lerman, K. (2010).

Predicting influential users in online social networks.
In *Proceedings of KDD workshop on Social Network Analysis (SNA-KDD)*.

Gong, N. Z., Talwalkar, A., Mackey, L., Huang, L., Shin, E. C. R., Stefanov, E., Song, D., et al. (2011).

Jointly predicting links and inferring attributes using a social-attribute network (san).
arXiv preprint arXiv:1112.3265.

Hakimi, S. L. (1964).

Optimum locations of switching centers and the absolute centers and medians of a graph.
Operations Research, 12(3):450–459.

Hirsch, J. E. (2005).

An index to quantify an individual's scientific research output.
Proceedings of the National Academy of Sciences of the United States of America, 102(46,):16569–16572.

References II

Jeh, G. and Widom, J. (2003).

Scaling personalized web search.
In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 271–279, New York, NY, USA. ACM Press.

Opsahl, T., Agneessens, F., , and Skvoretz, J. (2010).

Node centrality in weighted networks: Generalizing degree and shortest paths.
Social Networks, 32(3):245–251.

Sergey, B., Motwani, R., , Page, L., and Winograd, T. (1998).

What can you do with a web in your pocket?
IEEE Data Engineering Bulletin, 21:37–47.

Tang, J., Sun, J., , Wang, C., and Yang, Z. (2009).

Social influence analysis in large-scale networks.
In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 807–816, New York, NY, USA. ACM.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008).

Arnetminer: extraction and mining of academic social networks.
In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM.

Wasserman, S. and Faust, K. (1994).

Social network analysis: Methods and applications.
Cambridge Univ Pr.

Yin, Z., Gupta, M., Weninger, T., and Han, J. (2010).

A unified framework for link recommendation using random walks.
In *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, pages 152–159. IEEE.

References III

Zhou, Y., Cheng, H., and Yu, J. X. (2009).

Graph clustering based on structural/attribute similarities.

Proceedings of the VLDB Endowment, 2:718–729.