The Pennsylvania State University
The Graduate School

IDENTIFYING SIMILAR OBJECTS IN SOCIAL NETWORKS AND DIGITAL LIBRARIES

A Dissertation in
Computer Science and Engineering
by
Hung-Hsuan Chen

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2013

The dissertation of Hung-Hsuan Chen was reviewed and approved* by the following:

C. Lee Giles
David Reese Professor of Information Sciences and Technology
Dissertation Advisor, Chair of Committee

Robert Collins
Associate Professor of Computer Science and Engineering

Wang-Chien Lee
Associate Professor of Computer Science and Engineering

David J. Miller
Professor of Electrical Engineering

Lee Coraor
Director of Academic Affairs

*Signatures are on file in the Graduate School.

Abstract

With the rise of the computer age, various kinds of information can be easily accessed in digital format. However, the objects found within this information, such as people, places, dates, and firms, form a tangled and complex relationship that is usually challenging to untangle.

In this dissertation, we aim to unravel the relationship among objects to the finest extent: what are the similarity levels between any pairs of objects. Discovering similar objects can be the foundation of several research problems and applications. For example, objects can be clustered into several groups by merging similar objects together. This merging process can be recursively performed such that a hierarchical structure of these terms is constructed. In addition, the hidden relationship among objects can be inferred by examining the similar objects that do not explicitly interact with each other.

This dissertation examines the problem of discovering similar objects in two different settings: (1) discovering similar objects based on the interaction among them, and (2) discovering similar objects based on their meta-data. We will mainly focus on the first setting. The interactions among objects are modeled by a network structure, in which each node represents one object, and an edge is presented if the two objects have interacted with each other. In the second setting, we examine the similarity problem where additional information other than interacting history is available. In the second setting, we targeted digital library objects, such as papers, authors, published venues (i.e., the published conference or journal), etc. The meta-data of these objects could be, for example, the citation counts of the paper, the affiliation of the author, and the topics of the conference. These meta-data are utilized to infer the similar objects, such as similar terms, similar venues, or relevant authors given a topic.

To validate our proposed models and methodologies, we conducted various experiments on several different data sets to discover the hidden relationship among

the target objects. This includes (1) the relationship between the authors, papers, and venues in the given digital library, (2) the actors, actresses, and the movies in the given movie information, and (3) the diseases and the genes of patients. In addition, we implemented two live systems based on CiteSeerX digital library to bring several of these research results into practical products. The first system, CollabSeer, recommends potential collaborators based on a user's research interest and previous coauthoring behaviors. The second one, CSSeer, recommends a list of experts given a term of interest based on the similarity score between the query term and the publication and citation history of the authors. Both systems are highly efficient in handling more than one million papers and over 300 thousand disambiguated authors.

Acknowledgments

Science educators usually present knowledge in carefully arranged chapters such that students can learn in a logical order. However, students usually overlook the hardworking behind the scientific discoveries. Even though the real story of a successful research usually accompanies with failures, stupid mistakes, and laborious work, the frustration process are not presented in the research papers or textbooks. As a result, the real efforts and the original derivation process are hidden in the carefully re-arranged and organized text. Even a simple and seem-to-be straightforward equation may be derived by numerous conjectures, experiments, and trial-and-errors.

I started to realize the challenging process after being a graduate student. Luckily, my academic adviser, Dr. C. Lee Giles, provides rich resources so that I can pay more attention on thinking and conducting experiments. More importantly, he creates a wonderful environment to encourage peer discussion. Through these discussions, I could better see the strength and the weakness of my thoughts, and shape vague thoughts into more mature ideas. After discussion and imagination, Dr. Giles often proposes down-to-the-earth suggestions so that we may convert ideas into research papers and practical products. Even though research process can sometimes be frustrating, converting an idea into a product is always a joyful moment and encourages me to continue.

In addition to Dr. Giles, I would like to thank Dr. David Millers careful editing and insightful comments on my proposal. His suggestions inspired me to quantify the predictive power of young and old links in a network and challenge whether the commonly discovered network statistics still hold when links do decay over time.

I would like to thank Dr. Wang-Chien Lee for offering a great social networking course. Dr. Lee paid great efforts to encourage the discussions in class. This is one of the most rewarding courses I took at PSU.

I would like to thank several great researchers in my earlier student life. I would like to thank Dr. Jia-Shung Wang, who showed me the beauty of math and influenced me to go abroad for further study, Dr. Shou-De Lin, who helped

me to view several research problems more systematically, Dr. Kuan-Ta Chen, who shared several of his research methods and secret weapons with me. I also published my first first author paper with Dr. Chen.

Finally, I would like to thank my father in heaven. He was not a scientist, but in many daily works, he behaved and thought like a scientist. He was the first person to teach me induction, although he may not be aware of this term. In many ways, I am deeply influenced by him.

Bibliography

[1] Watts, D. and S. Strogatz (1998) "Collective dynamics of small-world networks," *Nature*, **393**(6684), pp. 440–442.

[2] Barabási, A. and R. Albert (1999) "Emergence of scaling in random networks," *Science*, **286**(5439), p. 509.

[3] Tang, J., J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su (2008) "ArnetMiner: extraction and mining of academic social networks," in *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 990–998.

[4] Friberg, S. (2012), "Evolution, Science, and Religion 17: On Heirarchies in Science and on Reductionism," .
URL http://www.commongroundgroup.net/2012/06/03/evolution-science-and-religion-17-on-heirarchies-of-science-and-reductionis

[5] Anderson, P. W. et al. (1972) "More is different," *Science*, **177**(4047), pp. 393–396.

[6] Zanin, M., E. Menasalvas, S. Boccaletti, and P. Sousa (2013) "Feature selection in the reconstruction of complex network representations of spectral data," *PloS one*, **8**(8), p. e72045.

[7] Erdös, P. and A. Rényi (1959) "On random graphs, I," *Publicationes Mathematicae (Debrecen)*, **6**, pp. 290–297.

[8] Chung, F. and L. Lu (2001) "The diameter of random sparse graphs," *Advances in Applied Math*, **26**(4), pp. 257–279.

[9] ——— (2002) "The average distances in random graphs with given expected degrees," *Proceedings of the National Academy of Sciences*, **99**(25), pp. 15879–15882.

[10] Leskovec, J., J. Kleinberg, and C. Faloutsos (2005) "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining*, ACM.

[11] Rapoport, A. (1953) "Spread of information through a population with socio-structural bias: I. Assumption of transitivity," *Bulletin of Mathematical Biology*, **15**(4), pp. 523–533.

[12] Nguyen, V., C. Leung, and E. Lim (2011) "Modeling link formation behaviors in dynamic social networks," *Social Computing, Behavioral-Cultural Modeling and Prediction*, pp. 349–357.

[13] Chen, H.-H., L. Gou, X. Zhang, and C. L. Giles (2011) "Capturing missing edges in social networks using vertex similarity," in *Proceedings of the 6th ACM International Conference on Knowledge Capture*, ACM, pp. 195–196.

[14] ——— (2011) "Collabseer: A search engine for collaboration discovery," in *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries*, ACM, pp. 231–240.

[15] Leskovec, J., L. Backstrom, R. Kumar, and A. Tomkins (2008) "Microscopic evolution of social networks," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 462–470.

[16] Romero, D. and J. Kleinberg (2010) "The directed closure process in hybrid social-information networks, with an analysis of link formation on Twitter," in *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pp. 138–145.

[17] Tan, P.-N., M. Steinbach, and V. Kumar (2006) *Introduction to data mining*, Pearson Addison Wesley Boston.

[18] Salton, G. (1989) *Automatic text processing: the transformation, analysis, and retrieval of information by computer*.

[19] Adamic, L. and E. Adar (2003) "Friends and neighbors on the web," *Social Networks*, **25**(3), pp. 211–230.

[20] Ravasz, E., A. Somera, D. Mongru, Z. Oltvai, and A. Barabási (2002) "Hierarchical organization of modularity in metabolic networks," *Science*, **297**(5586), p. 1551.

[21] Price, D. (1976) "A general theory of bibliometric and other cumulative advantage processes," *Journal of the American Society for Information Science*, **27**(5), pp. 292–306.

[22] Eisenberg, E. and E. Levanon (2003) "Preferential attachment in the protein network evolution," *Physical Review Letters*, **91**(13), p. 138701.

[23] Newman, M. (2001) "Clustering and preferential attachment in growing networks," *Physical Review E*, **64**(2), p. 25102.

[24] Zhou, T., L. Lü, and Y.-C. Zhang (2009) "Predicting missing links via local information," *The European Physical Journal B-Condensed Matter and Complex Systems*, **71**(4), pp. 623–630.

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

[26] Nowell, D. and J. Kleinberg (2003) "The link prediction problem for social networks," in *CIKM03: Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pp. 556–559.

[27] Leicht, E., P. Holme, and M. Newman (2006) "Vertex similarity in networks," *Physical Review E*, **73**(2), p. 26120.

[28] Jeh, G. and J. Widom (2002) "SimRank: A measure of structural-context similarity," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 538–543.

[29] Zhao, P., J. Han, and Y. Sun (2009) "P-Rank: a comprehensive structural similarity measure over information networks," in *Proceeding of the 18th ACM Conference on Information and Knowledge Management*, ACM, pp. 553–562.

[30] Xi, W., E. Fox, W. Fan, B. Zhang, Z. Chen, J. Yan, and D. Zhuang (2005) "SimFusion: measuring similarity using unified relationship matrix," in *Proceedings of the 28th Annual International SIGIR Conference on Research and Development in Information Retrieval*, ACM.

[31] Cai, Y., M. Zhang, C. Ding, and S. Chakravarthy (2010) "Closed form solution of similarity algorithms," in *Proceedings of the 33rd International SIGIR Conference on Research and Development in Information Retrieval*, ACM.

[32] Acar, E., D. Dunlavy, and T. Kolda (2009) "Link prediction on evolving data using matrix and tensor factorizations," in *IEEE International Conference on Data Mining Workshops*, IEEE, pp. 262–269.

[33] Li, C., J. Han, G. He, X. Jin, Y. Sun, Y. Yu, and T. Wu (2010) "Fast computation of simrank for static and dynamic information networks," in *Proceedings of the 13th International Conference on Extending Database Technology*, ACM, pp. 465–476.

[34] He, G., H. Feng, C. Li, and H. Chen (2010) "Parallel SimRank computation on large graphs with iterative aggregation," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM.

[35] Yu, W., X. Lin, W. Zhang, Y. Zhang, and J. Le (2012) "SimFusion+: extending simfusion towards efficient estimation on large and dynamic networks," in *Proceedings of the 35th International SIGIR Conference on Research and Development in Information Retrieval*, ACM.

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

[37] Roth, M., A. Ben-David, D. Deutscher, G. Flysher, I. Horn, A. Leichtberg, N. Leiser, Y. Matias, and R. Merom (2010) "Suggesting friends using the implicit social graph," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 233–242.

[38] Al Hasan, M., V. Chaoji, S. Salem, and M. Zaki (2006) "Link prediction using supervised learning," in *Workshop on Link Analysis, Counter-terrorism and Security*.

[39] Wang, C., V. Satuluri, and S. Parthasarathy (2007) "Local probabilistic models for link prediction," in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, IEEE, pp. 322–331.

[40] Cukierski, W., B. Hamner, and B. Yang (2011) "Graph-based features for supervised link prediction," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, IEEE, pp. 1237–1244.

[41] Coyle, K. (2006) "Mass digitization of books," *The Journal of Academic Librarianship*, **32**(6), pp. 641–645.

[42] Jones, S. and G. Paynter (2002) "Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications," *Journal of the American Society for Information Science and Technology*, **53**(8), pp. 653–677.

[43] Nguyen, T. and M. Kan (2007) "Keyphrase extraction in scientific publications," *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, pp. 317–326.

[44] Vincent, L. (2007) "Google Book Search: Document understanding on a massive scale," in *Document Analysis and Recognition, 2007. ICDAR 2007. 9th International Conference on*, vol. 2, IEEE, pp. 819–823.

[45] Witten, I., G. Paynter, E. Frank, C. Gutwin, and C. Nevill-Manning (1999) "KEA: Practical automatic keyphrase extraction," in *Proceedings of the Fourth ACM Conference on Digital Libraries*, ACM, pp. 254–255.

[46] Kim, S. and M. Kan (2009) "Re-examining automatic keyphrase extraction approaches in scientific articles," in *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, Association for Computational Linguistics, pp. 9–16.

[47] Treeratpituk, P., P. Teregowda, J. Huang, and C. L. Giles (2010) "SEERLAB: A system for extracting key phrases from scholarly documents," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, pp. 182–185.

[48] Grineva, M., M. Grinev, and D. Lizorkin (2009) "Extracting key terms from noisy and multitheme documents," in *Proceedings of the 18th International Conference on World wide Web*, ACM, pp. 661–670.

[49] Mihalcea, R. and A. Csomai (2007) "Wikify!: linking documents to encyclopedic knowledge," in *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management*, ACM, pp. 233–242.

[50] Pedersen, T., S. Patwardhan, and J. Michelizzi (2004) "WordNet:: Similarity: measuring the relatedness of concepts," in *Demonstration Papers at HLT-NAACL 2004*, Association for Computational Linguistics, pp. 38–41.

[51] Ruppenhofer, J., M. Ellsworth, M. Petruck, C. Johnson, and J. Scheffczyk (2006), "Framenet ii: Extended theory and practice," .

[52] Turney, P. (2001) "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL," .

[53] Zhou, D. X. and P. Resnick (2009) "Assessment of conversation comentions as a resource for software module recommendation," in *Proceedings of the Third ACM Conference on Recommender Systems*, ACM, pp. 133–140.

[54] Strube, M. and S. Ponzetto (2006) "WikiRelate! Computing semantic relatedness using Wikipedia," in *Proceedings of the National Conference on Artificial Intelligence*, vol. 21, AAAI Press, p. 1419.

[55] Gabrilovich, E. and S. Markovitch (2007) "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pp. 1606–1611.

[56] Witten, I. and D. Milne (2008) "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links," in *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pp. 25–30.

[57] Milne, D. and I. Witten (2008) "Learning to link with wikipedia," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ACM, pp. 509–518.

[58] Zhu, J., B. Zhang, Z. Nie, J. Wen, and H. Hon (2007) "Webpage understanding: an integrated approach," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM.

[59] Balog, K., L. Azzopardi, and M. De Rijke (2006) "Formal models for expert finding in enterprise corpora," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 43–50.

[60] Fang, H. and C. Zhai (2007) "Probabilistic models for expert finding," *Advances in Information Retrieval*.

[61] Deng, H., I. King, and M. Lyu (2008) "Formal models for expert finding on DBLP bibliography data," in *Data Mining, 2008. ICDM'08. 8th IEEE International Conference on*, IEEE, pp. 163–172.

[62] Campbell, C., P. Maglio, A. Cozzi, and B. Dom (2003) "Expertise identification using email communications," in *Proceedings of the 12th International Conference on Information and Knowledge Management*, ACM, pp. 528–531.

[63] Deng, H., J. Han, B. Zhao, Y. Yu, and C. Lin (2011) "Probabilistic topic models with biased propagation on heterogeneous information networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 1271–1279.

[64] Li, J., J. Tang, J. Zhang, Q. Luo, Y. Liu, and M. Hong (2007) "Eos: expertise oriented search using social networks," in *Proceedings of the 16th International Conference on World Wide Web*, ACM, pp. 1271–1272.

[65] Zhang, J., J. Tang, and J. Li (2007) "Expert finding in a social network," *Advances in Databases: Concepts, Systems and Applications*, pp. 1066–1069.

[66] Fan, W., X. Wang, and Y. Wu (2013) "ExpFinder: Finding experts by graph pattern matching," in *Data Engineering (ICDE), 2013, IEEE 29th International Conference on*, IEEE, pp. 1316–1319.

[67] Deng, H., J. Han, M. Lyu, and I. King (2012) "Modeling and exploiting heterogeneous bibliographic networks for expertise ranking," in *Proceedings of the 12th ACM/IEEE Joint Conference on Digital Libraries*, ACM.

[68] Gollapalli, S., P. Mitra, and C. Giles (2011) "Ranking authors in digital libraries," in *Proceedings of the 11th ACM/IEEE Joint Conference on Digital libraries*, ACM.

[69] Bian, J., Y. Liu, D. Zhou, E. Agichtein, and H. Zha (2009) "Learning to recognize reliable users and content in social media with coupled mutual reinforcement," .

[70] Albert, R. and A. Barabási (2002) "Statistical mechanics of complex networks," *Reviews of Modern Physics*, **74**(1), pp. 47–97.

[71] Adafre, S. and M. de Rijke (2005) "Discovering missing links in Wikipedia," in *Proceedings of the 3rd International Workshop on Link Discovery*, ACM, pp. 90–97.

[72] Bilenko, M., R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg (2005) "Adaptive name matching in information integration," *Intelligent Systems, IEEE*, **18**(5), pp. 16–23.

[73] Clark, P., J. Thompson, K. Barker, B. Porter, V. Chaudhri, A. Rodriguez, J. Thoméré, S. Mishra, Y. Gil, P. Hayes, et al. (2001) "Knowledge entry as the graphical assembly of components," in *Proceedings of the 1st International Conference on Knowledge Capture*, ACM, p. 29.

[74] Chen, H.-H., L. Gou, X. L. Zhang, and C. L. Giles (2013) "Towards the Discovery of Diseases Related by Genes Using Vertex Similarity Measures," in *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*, IEEE, pp. 505–510.

[75] Migliore, M., V. Martorana, and F. Sciortino (1990) "An algorithm to find all paths between two nodes in a graph," *Journal of Computational Physics*, **87**(1), pp. 231–236.

[76] Desrosiers, C. and G. Karypis (2010) "Enhancing link-based similarity through the use of non-numerical labels and prior information," in *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, ACM, pp. 26–33.

[77] Goh, K., M. Cusick, D. Valle, B. Childs, M. Vidal, and A. Barabási (2007) "The human disease network," *Proceedings of the National Academy of Sciences*, **104**(21), p. 8685.

[78] Chen, H.-H., L. Gou, X. Zhang, , and C. L. Giles (2012) "Discovering missing links in networks using vertex similarity measures," in *The 27th ACM Symposium on Applied Computing*, ACM.

[79] Treeratpituk, P. and C. L. Giles (2009) "Disambiguating authors in academic publications using random forests," in *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM, pp. 39–48.

[80] Chen, H.-H., D. J. Miller, and C. L. Giles (2013) "The predictive value of young and old links in a social network," in *Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks*, ACM, pp. 43–48.

[81] Ley, M. and P. Reuther (2006) "Maintaining an online bibliographical database: The problem of data quality," *Extraction et gestion des connaissances (EGC'2006), Actes des sixiémes journées Extraction et Gestion des Connaissances, Lille, France*, pp. 17–20.

[82] Breiman, L. (1996) "Bagging predictors," *Machine learning*, **24**(2), pp. 123–140.

[83] Schapire, R. (2001) "The boosting approach to machine learning: an overview," .

[84] Ugander, J., B. Karrer, L. Backstrom, and C. Marlow (2011) "The anatomy of the Facebook social graph," *arXiv Preprint arXiv:1111.4503*.

[85] Dunbar, R. (1993) "Coevolution of neocortical size, group size and language in humans," *Behavioral and Brain Sciences*, **16**(4).

[86] Kempe, D., J. Kleinberg, and É. Tardos (2003) "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 137–146.

[87] Leskovec, J., A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance (2007) "Cost-effective outbreak detection in networks," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 420–429.

[88] Fan, W., J. Li, J. Luo, Z. Tan, X. Wang, and Y. Wu (2011) "Incremental graph pattern matching," in *Proceedings of the 2011 International Conference on Management of Data, SIGMOD*, vol. 11, pp. 925–936.

[89] O'Doherty, D., S. Jouili, and P. Van Roy (2012) "Towards trust inference from bipartite social networks," in *Proceedings of the 2nd ACM SIGMOD Workshop on Databases and Social Networks*, ACM, pp. 13–18.

[90] Chen, H.-H., Y.-B. Ciou, and S.-D. Lin (2012) "Information propagation game: a tool to acquire humanplaying data for multiplayer influence maximization on social networks," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 1524–1527.

[91] Chen, H.-H. and C. L. Giles (2013) "ASCOS: an asymmetric network structure context similarity measure," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ACM, pp. 442–449.

[92] Panigrahy, R., M. Najork, and Y. Xie (2012) "How user behavior is related to social affinity," in *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM)*, ACM.

[93] Maguitman, A., F. Menczer, F. Erdinc, H. Roinestad, and A. Vespignani (2006) "Algorithmic computation and approximation of semantic similarity," *World Wide Web*, **9**(4), pp. 431–456.

[94] Tversky, A. (1977) "Features of similarity." *Psychological Review*, **84**(4), p. 327.

[95] Antonellis, I., H. G. Molina, and C. C. Chang (2008) "Simrank++: query rewriting through link analysis of the click graph," *Proceedings of the VLDB Endowment*, **1**(1), pp. 408–421.

[96] Lizorkin, D., P. Velikhov, M. Grinev, and D. Turdakov (2008) "Accuracy estimate and optimization techniques for simrank computation," *Proceedings of the VLDB Endowment*, **1**(1), pp. 422–433.

[97] Cullum, J. and R. Willoughby (2002) *Lanczos algorithms for large symmetric eigenvalue computations*, vol. 1, Society for Industrial Mathematics.

[98] Golub, G. and C. Van Loan (1996) *Matrix computations*, Johns Hopkins Univ Pr.

[99] Adams, M. (2001) "A distributed memory unstructured Gauss-Seidel algorithm for multigrid smoothers," in *Supercomputing, ACM/IEEE 2001 Conference*, IEEE.

[100] Knuth, D. (1993) *The Stanford GraphBase: a platform for combinatorial computing*, ACM Press.

[101] Nelson, D. L., C. L. McEvoy, and T. A. Schreiber (2004) "The University of South Florida free association, rhyme, and word fragment norms," *Behavior Research Methods, Instruments, & Computers*, **36**(3), pp. 402–407.

[102] Lichtnwalter, R. and N. V. Chawla (2012) "Link prediction: fair and effective evaluation," in *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, IEEE, pp. 376–383.

[103] Zhang, J., C. Wang, P. S. Yu, and J. Wang (2013) "Learning latent friendship propagation networks with interest awareness for link prediction," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, ACM, pp. 63–72.

[104] Chen, H.-H., P. Treeratpituk, P. Mitra, and C. L. Giles (2013) "CSSeer: an expert recommendation system based on CiteseerX," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, ACM, pp. 381–382.

[105] Conry, D., Y. Koren, and N. Ramakrishnan (2009) "Recommender systems for the conference paper assignment problem," in *Proceedings of the Third ACM Conference on Recommender Systems*, ACM, pp. 357–360.

[106] Yimam-Seid, D. and A. Kobsa (2003) "Expert-finding systems for organizations: Problem and domain analysis and the DEMOIR approach," *Journal of Organizational Computing and Electronic Commerce*, **13**(1), pp. 1–24.

[107] Zobel, J. and A. Moffat (2006) "Inverted files for text search engines," *ACM Computing Surveys (CSUR)*, **38**(2), p. 6.

[108] Councill, I., C. Giles, and M. Kan (2008) "ParsCit: An open-source CRF reference string parsing package," in *Proceedings of LREC*, vol. 2008, European Language Resources Association (ELRA), pp. 661–667.

[109] Järvelin, K. and J. Kekäläinen (2002) "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems (TOIS)*, **20**(4), pp. 422–446.

[110] Chapelle, O., D. Metlzer, Y. Zhang, and P. Grinspan (2009) "Expected reciprocal rank for graded relevance," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ACM, pp. 621–630.

[111] Yang, Z., J. Tang, B. Wang, J. Guo, J. Li, and S. Chen (2009) "Expert2bole: From expert finding to bole search," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,(KDD'09)*, pp. 1–4.

[112] Kim, J., S.-K. Kim, and H. Yu (2013) "Scalable and parallelizable processing of influence maximization for large-scale social networks," in *Data Engineering (ICDE), 2013, IEEE 29th International Conference on*, IEEE, pp. 266–277.

[113] Dickens, L., I. Molloy, J. Lobo, P.-C. Cheng, and A. Russo (2012) "Learning Stochastic Models of Information Flow," in *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, IEEE, pp. 570–581.

[114] Lappas, T., K. Liu, and E. Terzi (2009) "Finding a team of experts in social networks," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 467–476.