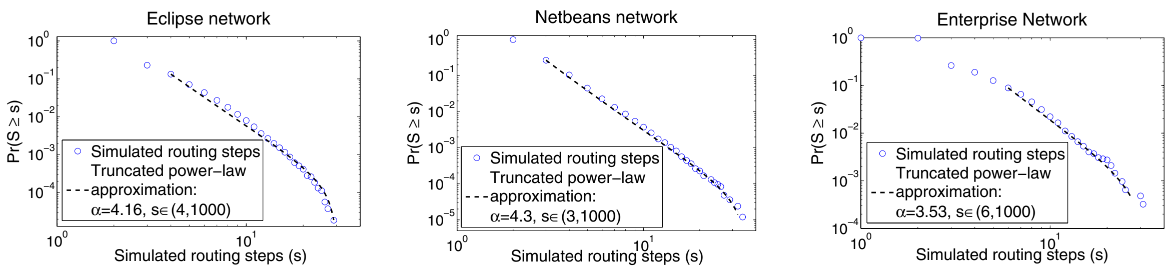

Figure 10: Simulated routing step distributions.

space, they are likely to be close to each other in the embedded space. The closeness between two nodes can be defined by the number of task transfers between them: the more frequent the task transfer, the closer the two nodes.

Figure 9 shows the two-dimensional embedding of the Netbeans network, using the spectral embedding method. The embedding can be regarded as a non-uniform distribution of nodes in an expertise space. Given the embedding, we can assign two-dimensional coordinates (x_1, x_2) to each node in the network, which enables distance measurement between pairs of nodes, a required input to the SGR model. Because we know the initiator and the resolver of each task, we then apply the SGR model to simulate the path of each task routing. The routing step distributions of the simulation for all three networks are shown in Figure 10. The simulated results match the observations well, as Figure 3 shows.

5.3 Combining the Two Models: A Case Study

Our network model simulates the static connectivity of a collaborative network, whereas our SGR model simulates the dynamic user behavior in information routing in a collaborative network. Combined together, these two models provide an unprecedented means of studying existing collaborative networks. It is important to study how the structure changes of a collaborative network can affect the efficiency of task execution, without changing the real-world network structure. This case study demonstrates the simulation method for our network and information routing models.

The case study is the problem management organization of a large IT service provider. To accommodate the evolving workload and human resources, the service provider needs to restructure the service agent network to deliver the optimal performance in resolving the problems reported by its clients. Currently, these restructuring decisions are made manually by experienced managers or consultants, without quantitative analysis as to how the resulting network will perform after the restructuring.

Our models can be used to provide analytical insights to the decision makers. First, one can use our network model to generate new network topologies with different structural constraints that need to be imposed in practice. Then, given a set of tasks, the efficiency of different networks can be evaluated through the task routing simulation guided by the SGR model. Here, we assume that a collaborative network of 5,000 service agents needs to be restructured. These agents are divided into K pools (expertise domains) based on their expertise. A important question is: "How does one select the optimal number K of pools, to provide the best efficiency in task execution?" Intuitively, a smaller value of K indicates that the agents are more generalized in their domain expertise, whereas a larger value of K suggests that the agents are more specialized in their domain expertise. Furthermore, with more domains, a task is less likely to be initially assigned to the right agent pool, which might lead to longer routing paths, because intra-domain routing is more likely to occur than inter-domain routing.

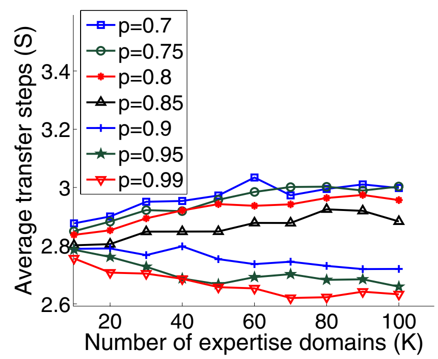

Figure 11: Evaluating the network structures.

For our analysis, we generate 10 collaborative networks, with 10 to 100 domains. In each network configuration, we simulate the routing of the same set of 100,000 tasks. The probability p of correctly assigning the task to the right domain is also taken into account in the simulation. For each task, first we select the resolver node with probability proportional to its incoming degree. Then, with probability p, the task is initiated within the same domain as the resolver; otherwise, the initiator is selected from outside the resolver's domain. We vary the "correct assignment probability" p from 0.7 to 0.99. For each value of p, we route the entire set of tasks in the 10 networks. The results of all simulations are shown in Figure 11. The y-axis shows the average number of transfer steps to the resolver for the entire set of tasks. Each curve shows the routing simulation results for a particular choice of p. Obviously, a lower average number of steps indicates a higher routing efficiency, because it usually takes less time when the tasks are routed to the resolver in fewer steps. As shown in the figure, when more tasks are initially assigned to the right domain, increasing the number of domains leads to better performance. When fewer tasks are initially assigned to the right domain, a smaller number of domains is more favorable.

Achieving a certain value of p, given various numbers of agent pools, has different implications in terms of training the initial assigner of the task: for the same value of p, the training cost typically increases as the number of agent pools increases, because the assigner must have stronger knowledge in matching the task with the correct expertise domain. Configuring the collaborative network into different numbers of expertise domains also has implications on the training cost for the agents. Given these implications, the decision maker can use our method to select the optimal number of agent pools that suits the enterprise's budget or other constraints.

6. RELATED WORK

Previous studies related to our work mainly belong to two categories: those that focus on network generation models, and those that analyze information flows in networks.

Network generation models. Generating synthetic networks that reflect statistics similar to real social networks has been of great interest to researchers in various fields. The Erdös-Rényi random network [9] is a classic random network, where any two nodes are connected according to a fixed probability. A regular lattice network is created with nodes placed on one or more dimensional lattices, *i.e.*, circle or grid, and each node is connected to its n nearest neighbors. Watts and Strogatz [32] added random rewiring to the regular lattice network such that the generated network has a small diameter as observed in a sample of the real social network [29]. Barabasi *et al.* [5] focused on the fact that many complex networks have degrees that follow a heavy-tail distribution and captured this phenomena by incrementally creating a random network, with new edges preferentially attached to already well-connected nodes. To comply with both the small-world effect and the power-law degree distribution, Makowiec [16] and Ree [21] proposed rewiring processes in a constant-size network based on the preferential attachment principle. Serrano *et al.* [26] developed a network generation model to reproduce self-similarity and scale invariance properties observed in real complex networks, by utilizing a hidden metric space with distance measurements. Sala *et al.* [24] studied how well the generated graphs match real social graphs extracted from Facebook.

Different from the existing graph generation models, our method contributes toward understanding how links are established and how members with different expertise interact with each other in real collaborative networks. Both the expertise awareness and expertise exposure of each member are taken into consideration in our model. It not only generates a network topology with statistical characteristics similar to real-world collaborative networks, but also can be seamlessly combined with our routing model to simulate human dynamics in these networks.

Information flow analysis. The spreading of information has been extensively studied under different network settings, *e.g.*, social networks, especially the World Wide Web, the e-mail network, biological networks, *etc.* Examples include the spread of innovations [12, 22, 27, 30], opinions, rumors and gossip [10, 11, 17], computer/biological viruses [15, 25] and marketing [8, 13]. More recently, Wang *et al.* [31] have studied how information propagates from person to person using e-mail forwarding, and Wu *et al.* [33] analyzed the information spreading pattern on Twitter. This type of information flow aims to reach and influence more people and, hence, to achieve a large impact. Most of the work has focused on analyzing patterns of the information spreading process. Kempe *et al.* [13] have addressed the question of how to choose a subset of nodes to initiate information spreading to maximize influence in a network.

In our work, we focus on another type of information flow: task-driven information flow, where the goal is to reach a user who can accomplish a task with a minimal number of transfer steps. Related to our problem, Milgram [19] demonstrated that short paths exist between any pair of nodes in a social network (*a.k.a.*, the small world phenomena). Kleinberg [14] investigated why decentralized navigation is efficient using a synthetic network lattice. Boguna *et al.* [6] studied the navigability of complex networks by running a greedy routing algorithm on synthetic networks generated by a model described in [26]. In the collaborative networks we studied, we observe that these networks exhibit degree distributions quite different from commonly-studied complex networks. Furthermore, the simple greedy algorithm does not provide a good approximation of information flow dynamics in collaborative networks. Thus, we developed the SGR model to evaluate the efficiency of task-driven information flow in such networks.

7. CONCLUSIONS

This study examined a special type of social networks – collaborative networks. Detailed observations of three real-world collaborative networks were presented along with the static network topology and dynamic information routing for each network. The collaborative networks exhibit not only the truncated power-law node degree distributions but also organizational constraints. Information routing in collaborative networks is different from routing in conventional complex networks, such as computer networks and airline networks, because of the random factors in human decision making. The routing steps in collaborative networks also follow a truncated power-law distribution, which implies that a considerable number of tasks travel along long sequences of steps before they are completed. Our results and observations for several independent sources are consistent with each other, and can be generalized to other real-world collaborative networks. They help in understanding the complicated behavior in human collaboration.

Based on real-world data, we developed a graph model to generate networks similar to real collaborative networks, and a stochastic routing algorithm to simulate the human dynamics of collaboration. The models are independently validated using real-world data. We demonstrated that the two models can be used to answer real-world questions, such as: *"How can one design a collaborative network to achieve higher efficiency?"* To the best of our knowledge, our work is the first attempt to understand human dynamics in collaborative networks and to estimate analytically the efficiency of real collaborative networks.

Acknowledgment

The first author, Gengxin Miao, was supported by an IBM Ph.D. Fellowship; she spent the Summer of 2011 at IBM T.J. Watson Research Center, where she conducted this research. This research was sponsored in part by the U.S. National Science Foundation under grant IIS-0917228 and by the Army Research Laboratory under cooperative agreement W911NF-09-2-0053 (NS-CTA). The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein.

8. REFERENCES

[1] Bugzilla: http://www.bugzilla.org/.
[2] Eclipse: http://www.eclipse.org/.
[3] Mozilla: http://www.mozilla.org/.
[4] Msr 2011 challenge: http://2011.msrconf.org/msr-challenge.html.
[5] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
[6] M. Boguna, D. Krioukov, and K. C. Claffy. Navigability of complex networks. *Nature Physics*, 5(1):74–80, 2008.
[7] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 2009.
[8] P. Domingos and M. Richardson. Mining the network value of customers. In *SIGKDD*, pages 57–66, 2001.
[9] P. Erdös and A. Rényi. On random graphs I. *Publicationes Mathematicae*, 6:290–297, 1959.
[10] S. Galam. Minority opinion spreading in random geometry. *The European Physical Journal B - Condensed Matter and Complex Systems*, 25(4):403–406, 2002.
[11] S. Galam. Modelling rumors: The no plane Pentagon French hoax case. *Physica A: Statistical Mechanics and Its Applications*, 320:571–580, 2003.
[12] X. Guardiola, A. Diaz-Guilera, C. J. Perez, A. Arenas, and M. Llas. Modeling diffusion of innovations in a social network. *Phys. Rev. E*, 66:026121, 2002.
[13] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *SIGKDD*, pages 137–146, 2003.
[14] J. Kleinberg. Small-world phenomena and the dynamics of information. In *NIPS*, page 2001. MIT Press, 2001.
[15] A. Lloyd and R. May. How viruses spread among computers and people. *Science*, 292(5520):1316–1317, 2001.
[16] D. Makowiec. Evolving network - simulation study. *The European Physical Journal B - Condensed Matter and Complex Systems*, 48:547–555, 2005.
[17] K. Malarz, Z. Szvetelszky, B. Szekf, and K. Kulakowski. Gossip in random networks. *ACTA Physica Polonica B*, 37, Nov. 2006.
[18] G. Miao, L. E. Moser, X. Yan, S. Tao, Y. Chen, and N. Anerousis. Generative models for ticket resolution in expert networks. In *KDD*, pages 733–742, 2010.
[19] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.
[20] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 2nd edition, Oct. 1992.
[21] S. Ree. Power-law distributions from additive preferential redistributions. *Phys. Rev. E*, 73:026115, Feb. 2006.
[22] E. M. Rogers. *Diffusion of Innovations*. Free Press, 4th edition, 1995.
[23] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.
[24] A. Sala, L. Cao, C. Wilson, R. Zablit, H. Zheng, and B. Y. Zhao. Measurement-calibrated graph models for social network experiments. In *WWW*, pages 861–870, 2010.
[25] R. P. Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev.*, 86(14):3200–3203, 2001.
[26] M. Serrano, D. Krioukov, and M. Boguna. Self-similarity of complex networks and hidden metric spaces. *Phys. Rev.*, 078701, 2008.
[27] D. Strang and S. A. Soule. Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual Review of Sociology*, 24(1):265–290, 1998.
[28] R. Taylor. *Constrained switchings in graphs*. Research report. University of Melbourne, Department of Mathematics, 1980.
[29] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.
[30] T. W. Valente. Network models of the diffusion of innovations. *Computational and Mathematical Organization Theory*, 2:163–164, 1996.
[31] D. Wang, Z. Wen, H. Tong, C. Y. Lin, C. Song, and A. L. Barabási. Information spreading in context. In *WWW*, pages 735–744, 2011.
[32] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.
[33] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on Twitter. In *WWW*, pages 705–714, 2011.