Managing Uncertainty in Social Networks

Eytan Adar and Christopher Ré
Department of Computer Science and Engineering
University of Washington
{eadar,chrisre}@cs.washington.edu

Abstract

Social network analysis (SNA) has become a mature scientific field over the last 50 years and is now an area with massive commercial appeal and renewed research interest. In this paper, we argue that new methods for collecting social nework strucuture, and the shift in scale of these networks, introduces a greater degree of imprecision that requires rethinking on how SNA techniques can be applied. We discuss a new area in data management, probabilistic databases, whose main research goal is to provide tools to manage and manipulate imprecise or uncertain data. We outline the application building blocks necessary to build a large scale social networking application and the extent to which current research in probabilisitc databases addresses these challenges.

1 Introduction

Though the field of Social Network Analysis (SNA) has developed over the past 50 or more years [21, 56], it is with the recent emergence of large-scale social networking studies and applications that techniques from this area have received a great deal of public attention. Because the data encapsulated by these networks provides the owners of a system with a mineable resource for marketing, health, communication, and other applications, commercial developers have rushed to construct *social network applications*. Such systems generally enable individuals to connect with old friends and colleagues and form bridges to new individuals in areas ranging from business (e.g. Visible Path [45] and Linked In [30]) to socialization (e.g. Facebook [19] and MySpace [42]) and to entertainment (e.g. iLike [11]). However, translating the research techniques of SNA to large scale applications is a daunting task. With large scale comes imprecision as applications depend on a new set of measurement instruments to collect their data and developers can no longer be completely confident that data about individuals, or the connections between them, is accurate. For example, data collected through automated sensors [9], anonymized communication data (e.g. e-mail headers [1]), and self-reporting/logging on Internet-scale networks [12, 23] as a proxy for real relationships and interactions causes some uncertainty. Furthermore, approximation algorithms [58] intended to calculate network properties (e.g. various centrality measures) on these increasingly large networks creates additional uncertainty. Traditionally, managing large scale datasets has been the domain of data management research and technologies which have almost always assumed that data is precise. In this paper we argue that the transition from research projects to commercial applications creates a need for tools that are able to support SNA techniques and that a critical component is the ability to

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

manage large scale imprecision. Specifically, we make the argument that SNA data can be modeled, managed, and mined effectively by emerging Probabilistic Databases (PDBs). Our discussion does not offer any new implementations or algorithms for PDBs but demonstrates if, when, and how PDBs can be leveraged in the context of SNA and related applications.

The starting point of SNA techniques is a graphical representation in which nodes—called *actors* in the SNA field—represent individuals or groups. An edge (potentially labeled) in this graphical view represents the relationship between actors and generally indicates the possibility of information flow between them. In the early history of SNA, this graph data was collected by survey, interview, and other observational techniques [21, 29, 56, 57]. While the results were potentially tainted by biased observations, missed observations, and misreporting, the intimate involvement of the researcher (frequently, over extended periods) provided some confidence that the data is precise. As those studying and utilizing social networks have moved to enormous scales, they have frequently sacrificed some accuracy as careful methodologies have become increasingly difficult or impossible. Furthermore, in wild and uncontrolled environments such as the Internet, biases can develop due to application design (e.g. default friends on MySpace) and malicious individuals (e.g. spammers building network connections in some automated way). The result of this "noise" is the introduction of tremendous levels of uncertainty in the data which are ill-supported by current large scale data management systems.

Probabilistic relational databases—the potential answer to these issues—have attracted renewed interest in the data management community [8, 14, 17, 22, 47, 52, 59]. A probabilistic database works much in the same was as a standard relational database, in which tuples (i.e. rows of data) can be stored, searched and aggregated in various ways using SQL. The defining characteristic of a probabilistic database is that to any tuple t, a probability is associated that indicates the probability t is in the database. While a standard relational database is intended to support a precise data model (e.g. Bob lives at "121 South Street"), a probabilistic database models uncertainty (e.g. there is 80% chance Bob lives at "121 South Street" and a 20% chance he lives at "50 West Street"). The motivating goal of this area is to provide application developers with the tools they need to *manage* imprecision while providing industrial scale performance. In this paper, we select a very simple data model called *tuple independence*, which is supported by all models in the recent literature. We refer the interested reader to work on more sophisticated models that are capable of representing any distribution (e.g. [47, 51, 52]) and research on models dealing with continuous data (e.g. sensor networks [8, 17]).

2 A Motivating Example

To understand the application of probabilistic databases (PDBs) in the context of social network research, we concentrate our efforts on a fictitious diffusion model[1] for music recommendations. Diffusion models are interesting in that they capture a range of application areas including epidemic models (e.g. [36, 40, 43]), innovation diffusion (e.g. [5, 10, 50, 54, 55]), and rumor and gossip propagation (e.g. [33]). Diffusion models are additionally relevant to both pure scientific discoveries about basic behavioral processes (e.g. [10, 25, 41]) and applied endeavors such as expert-finding networks [2], recommender systems [32], and public health community-building [5]. We later return to some of these applications by generalizing our example to a broader class of diffusion models.

Our system, graphically represented in Fig. 1, has two types of data about its users: standard actor/node information (e.g. name, age, residing city, etc.) (Fig. 1(b)) and preference data (e.g. music preferences) in terms of genre (Fig. 1(c)). Because we have determined the preference through a sampling methodology (e.g. by asking individuals to indicate their like or dislike of a set of songs), we are uncertain about its true value. This is modeled by assigning a probability to a tuple (e.g. (Kim,Country) is in Prefs with probability 0.75). To simplify our model, we assume that tuples that do not exist have a probability 0 (e.g. Alice does not like rap

[1] A full survey of this field is well beyond the scope of this paper. These citations represent interesting exemplars in this space. Some are early, influential publications, others represent more modern examples.

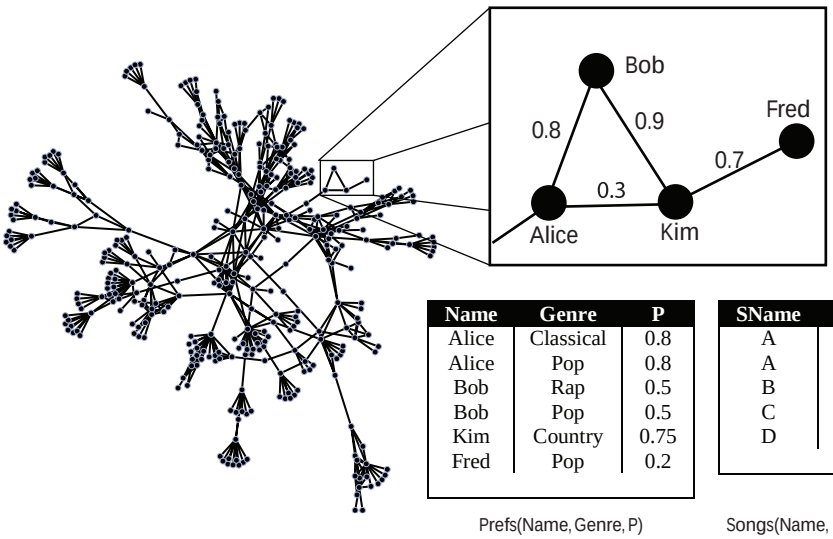

Figure 1a

Name	Age	City
Alice	25	Seattle
Fred	21	Reno
Bob	30	Seattle
Kim	27	LA

Users(Name, Age, City)

Figure 1b

Name	Genre	P
Alice	Classical	0.8
Alice	Pop	0.8
Bob	Rap	0.5
Bob	Pop	0.5
Kim	Country	0.75
Fred	Pop	0.2

Prefs(Name, Genre, P)

Figure 1c

SName	Genre	P
A	Country	0.1
A	Pop	0.5
B	Rap	0.5
C	Classical	0.8
D	Country	0.6

Songs(Name, Genre, Probability)

Figure 1d

Name1	Name2	P
Alice	Bob	0.8
Alice	Kim	0.3
Bob	Kim	0.9
Fred	Kim	0.7

MusicInfuence(Name1, Name2, P)

Figure 1e

Figure 1: Sample Data for Social Network Integration

```
SELECT U.name
FROM Users U, Prefs P, Songs S
WHERE U.name = P.name
  AND P.Genre = S.genre
  AND S.name = 'A'
```

(a)

```
SELECT U.Name
FROM Users U , Prefs P, Song S,
  MusicInfluence M, Recommends R
WHERE U.name = P.name AND P.Genre = S.genre
  AND M.name2 = U.name AND M.name1 = R.from
  AND R.To = U.name AND R.song = S.sname
  AND S.sname = 'D'
```

(b)

Figure 2: Sample Queries

music)[2]. Additionally, we have a table of songs (Fig. 1(d)). It is often not clear-cut to which genre a songs belongs, which we model by assigning a probability that a song belongs to a given genre. Given this data we can now ask questions of the form: for song A in Songs, "what is the probability that a user would like A?" This is expressed in SQL in Fig. 2(a). For example, the user Kim would have probability $0.75 * 0.1 = 0.075$ since we have assumed all tuples are independent (note that result probabilities are always returned alongside result tuple without the user needing to make an explicit request for those).

Consider an additional table, MusicInfluence, shown in Fig. 1(e), that describes a piece of our social network. To construct this table, we have assumed that users have explicitly defined their network (i.e. Alice has indicated that she is friends with Bob), and through some experience, we have assigned a probability on each edge indicating the likelihood that a musical recommendation from the first user will be picked up by the second (e.g. of the previous n recommendations, k were accepted. In our example, Alice influences Bob, with probability 0.8). We can use this table to write more interesting queries. Assuming that both Alice and Bob have recommended song D to Kim (Fig. 3a), we can ask, "what is the probability that Kim will be affected by these recommendations?" Intuitively, the SQL query Fig. 2(b)) is satisfied for Kim when, song D is country (0.6), Kim likes country (0.75) and either transmission from Alice or Bob is successful (0.8 and 0.9, resp.). Thus, Kim

[2][4] considers a semantic which is able to account for missing data as an "unknown" or "wildcard."

From	To	Song
Bob	Kim	D
Alice	Kim	D
…		

Recommends(From,To,Song)

(a) Recommends(from,to,song)

Name1	Name2	P
Alice	Bob	.8
Alice	Kim	.3
Bob	Fred	.1
…		

MusicSim(Name1,Name2,P)

(b) MusicSim(name1,name2;P)

Figure 3: Sample Data for Social Network Integration

is returned with a probability value of $0.6 * 0.75 * (1 - (1 - 0.8) * (1 - 0.9)) = 0.441$.

For the sake of exposition, we have simplified our example considerably; there are much more powerful types of models are possible: e.g. we could extended table to have the probability be given by type of music (e.g. (name1, name2, genre, probability)) or have more complicated correlations between the tuples (e.g. using *factors* [52] or BID tables [47]).

2.1 Application Building Blocks

In this section, we highlight some of the fundamental requirements of building a large scale SNA application and discuss the extent to which probabilistic databases can help address these requirements.

Data Analysis One of the most (if not the most) important business aspects of a social networking site is understanding the network. Put simply: you can not monetize what you do not understand. For example, in our scenario we may be interested in distributing free concert tickets for a new artist to a small subset of users. Since there is a cost to providing these tickets, we would like to find a small group of consumers who have a large amount of influence, i.e. a set of *trend-setters* or *influencers*. In the SNA world, this is a similar to a query for nodes with a high degree-centrality measure (e.g. the number of outgoing edges is high). Since the data are uncertain, one natural semantic is to rank users by their expected number of outgoing outgoing edges. These are essentially probabilistic *decision support* or OLAP style queries [7, 31]. Alternatively, we may only want to send the tickets to high value users, e.g. those with a high probability of having more than k edges, which has been consider in [48]. Although a large class of these queries can be handled efficiently by probabilistic databases, adapting more sophisticated SNA algorithms is an interesting direction for future research.

Scalability Large scale social network applications have very large datasets which need to be manipulated with good performance. Continuing our previous example, our webpage would need to issue queries such as "which users I am most similar to?" or "which users am I most similar to who live in Seattle?". This is daunting because the queries can combine several sources of probabilistic information. Sometimes, it is possible to correctly compute probabilistic database queries directly in SQL using a new technique called *safe plans* [14, 15, 46]. Intuitively, safe plans tell us when a probabilistic query can be computed by simply multiplying (and summing) probabilities. However, safe are not always possible, in which case a query may require approximation algorithms (e.g. a restricted kind of Monte Carlo simulation), which are slower but still tenable at large scales [47]. Often we are only interested in computing the top k answers, which can allow substantial improvements [47, 53]. Further, new research suggests that we may be able to materialize a probabilistic view which allows complex probabilistic datasets to scale even in to the tens and hundreds of gigabytes [49]. While probabilistic database research is still in its infancy, there are already techniques to scale up to huge datasets.

Physical and Semantic Independence In a large social network, as in any application with large numbers of users, tuning the backend is critical and requires constant tinkering. This effort is mitigated by a property that probabilistic databases on relational models inherit from relational databases, *physical independence*. PDBs achieve physical independence, because all interaction takes place through a query language that does not reference the physical layout on disk. Hence, data can be partitioned and indexed independently of how they are used by application code. Also important to sites that offer recommendations is the ability to compute, and propagate, qualitatively good recommendations. Thus, an approach is untenable if changing the code that computes the influence probability requires changing the code that displays the top ten most similar users. PDBs mitigates this problem because they achieve *semantic independence*. In particular, tuples have a clear probabilistic semantic independently of how they are computed. A tangible benefit is that we can decouple the computation of probabilities from their use in application code.

Maintainability A major problem in any large scale enterprise is maintaining, updating and debugging the data and applications built on it. As a concrete example, if the data, on which recommendations are based, changes (e.g. a user submits that they like new genre), the values in the relation should change as well. Also, if the end result of the computation breaks, how do we know how to fix it? There is very promising work in this direction based on *lineage* [51] in uncertain databases, which helps an administrator understand why or how a probability value is computed. We feel that the large body of work in the AI community on explaining a probabilistic proposition is a good starting place (e.g. [6, 35, 38], but one key remaining challenge is scaling these techniques to large datasets.

Integration Merging social networks is interesting from a research perspective as well as a business perspective [34]. For example, consider merging the network described above and an independent friendship network (e.g. Facebook). Intuitively, by leveraging more information the merging of two networks should provide higher answer quality and also allows us to ask queries not answerable by either network alone. For example, suppose we want to sell concert tickets for an intimate venue that only sells tickets in blocks of four (e.g. for tables). We would like to know, which users have three friends with similar tastes in music and live in the same area. To do this, we need to know both a persons friends and their taste in music, information not available in either networks by themselves. There are many difficult problems in integration, e.g. *entity resolution* or *reference reconciliation* [18, 20, 27, 61]. However, we believe a probabilistic databases provides a solid framework to model the uncertainty of inherent in the integration process.

Handling Missing Data While we may have an explicitly defined network it is always possible that we are missing certain important edges. This may be due to misreporting or flawed instrumentation but the outcome is an incomplete network. The idea that edges can be inferred has been studied extensively (e.g. by [24]). For our example we may use a simple algorithm that calculates the pairwise similarity between individuals based on the musical preference (and potentially their neighborhood). We model the output of the matching procedure using a probabilistic relation, which we call MusicSim. A snippet of data is shown in Fig. 3(b). A tuple in MusicSim means that name1 and name2 are similar, with probability given by the P attribute. For example, we might find that Alice and Bob have similar music tastes with probability 0.8, but Alice and Kim have similar music tastes with only probability 0.3. This is a powerful idea that is simplified greatly by the use of a probabilistic database.

2.2 Beyond Music

Though we have concentrated on a specific type of diffusion network above, there are clearly many application areas beyond music recommendation. An epidemic model, for example, may take into account susceptibility to a certain disease based on individual features, transmission probabilities assuming repeated contacts, probabilities

of immunization and other complex dynamics ([44]). A corporate network may take into account hierarchical and managerial influence on adoption of innovations. Clearly, correctly generating such models is a difficult and time consuming task, but managing and querying this type of imprecise information—especially in large scales—may be aided by the use of a probabilistic database.

3 Additional Application Areas

There are many additional areas in which social network analysis and applications are starting to be utilized in which the data is inherently imprecise. We select two of these areas that we think present particular important and interesting and where probabilistic databases have already received some attention.

Privacy and Anonymization As the use of social network information becomes more prevalent, it is important to recognize the privacy concerns of individuals. To understand the implications of social networks for privacy rights, a number of researchers [3, 26] have begun to explore how social networking data can, or can not, be anonymized using data perturbation techniques. We believe that probabilistic databases can play an interesting role in moving theoretical techniques of privacy-preservation (e.g. [13, 37, 39]) into large scale applications.

Homeland Security A sub-area of SNA that recently received a lot of attention is the analysis of terrorist networks. Here, SNA is focused on identifying "critical" individuals in the network. A report to congress by DARPA [16] about the now defunct Genisys program, highlighted the inadequacy of standard relational databases for the task and the need for "probabilistic database representing and dealing with uncertainty". Interestingly, the program was part of the TIA project that was defunded due to privacy concerns. However, according to media reports essentially the same program is still funded, under the name Topsail [28, 60].

4 Conclusion

In this paper we have argued that probabilistic databases are a useful paradigm for those who want to build social networking applications. The inherent imprecision and uncertainty of large-scale social network analysis, both in collection and analysis, does not need to add tremendous complexity to researchers and application designers. Even in their nascent state, probabilistic databases have much to offer social networking analysis and applications by handling the models, scaling, maintenance and analysis needs. Furthermore, we believe that social networks are an important motivating application for probabilistic database research. The growth of research and economic interest in social networking applications has generated a tremendous set of potential consumers of probabilistic databases. We have briefly discussed a number of interesting open research and technical problems to enable and support a wider range of social network applications. A mutually beneficial relationship between these two communities, especially during the rapid growth in both domains, will likely lead to many novel algorithms, techniques, and systems beyond anything we have imagined in this paper.

5 Acknowledgements

The authors would like to thank Bernie Hogan, Lada Adamic, Dan Weld, and Mike Cafarella for their comments and discussions. Eytan Adar is funded by an ARCS and NSF Graduate Fellowship.

References

[1] L. A. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187–203, 2005.

[2] E. Adar, R. Lukose, C. Sengupta, J. Tyler, and N. Good. Shock: A privacy-preserving knowledge network. *Information Systems Frontiers*, 5(1), 2003.
[3] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou R3579X? anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of WWW 2007*, 2007.
[4] D. Barbará, H. Garcia-Molina, and D. Porter. The management of probabilistic data. *IEEE Trans. Knowl. Data Eng.*, 4(5):487–502, 1992.
[5] D. M. Berwick. Disseminating innovations in health care. *Journal of the American Medical Association*, 289(15), 2003.
[6] A. Borgida, E. Franconi, and I. Horrocks. Explaining ALC subsumption. In Werner Horn, editor, *ECAI*, pages 209–213. IOS Press, 2000.
[7] Douglas Burdick, Prasad M. Deshpande, T. S. Jayram, Raghu Ramakrishnan, and Shivakumar Vaithyanathan. OLAP over uncertain and imprecise data. *VLDB J.*, 16(1):123–144, 2007.
[8] R. Cheng, D. Kalashnikov, and S. Prabhakar. Evaluating probabilistic queries over imprecise data. In *Proceedings of ACM SIGMOD Conference*, 2003.
[9] T. Choudhury, M. Philipose, D. Wyatt, and J. Lester. Towards activity databases: Using sensors and statistical models to summarize people's lives. *IEEE Data Eng. Bull.*, 29(1):49–58, 2006.
[10] J. Coleman, E. Katz, and H. Menzel. The diffusion of an innovation among physicians. *Sociometry*, 20(4), December 1957.
[11] Garage Band Corp. `www.ilike.com`.
[12] d. m. boyd. Friendster and publicly articulated social networking. In *Proceedings of CHI 2004*, pages 1279–1282, 2004.
[13] N. Dalvi, G. Miklau, and D. Suciu. Asymptotic conditional probabilities for conjunctive queries. In *ICDT*, 2005.
[14] N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. In *VLDB*, Toronto, Canada, 2004.
[15] N. Dalvi and D. Suciu. Management of probabilistic data: Foundations and challenges. In *PODS*, 2007.
[16] DARPA. Report to congress regarding the terrorism information awareness program, `http://www.epic.org/privacy/profiling/tia/may03_report.pdf`, 2003.
[17] A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, and W. Hong. Model-driven data acquisition in sensor networks, 2004.
[18] X. Dong, A. Y. Halevy, and J. Madhavan. Reference reconciliation in complex information spaces. In Fatma Özcan, editor, *SIGMOD Conference*, pages 85–96. ACM, 2005.
[19] Facebook. `www.facebook.com`.
[20] I. Felligi and A. Sunter. A theory for record linkage. *Journal of the American Statistical Society*, 64:1183–1210, 1969.
[21] L. C. Freeman. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press, 2004.
[22] A. Fuxman and R. J. Miller. First-order query rewriting for inconsistent databases. In *ICDT*, pages 337–351, 2005.
[23] L. Garton, C. Haythornthwaite, and B. Wellman. Studying online social networks. *Journal of Computer-Mediated Communication*, 3, 1997.
[24] L. Getoor and C. P. Diehl. Link mining: A survey. *SIGKDD Explorations*, 7(2), 2005.
[25] M. Granovetter. Threshold models of collective behavior. *The American Journal of Sociology*, 83(6), May 1978.
[26] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava. Anonymizing social networks. Technical Report 07-19, University of Massachusetts Amherst, CS Department, March 2007.
[27] M. A. Hernández and S. J. Stolfo. The merge/purge problem for large databases. In Michael J. Carey and Donovan A. Schneider, editors, *SIGMOD Conference*, pages 127–138. ACM Press, 1995.
[28] M. Hirsh. Wanted: Competent big brothers. *MSNBC*, `http://www.msnbc.msn.com/id/11238800/site/newsweek/`, Feb 2006.
[29] B. Hogan, J. A. Carrasco, and B. Wellman. Visualizing Personal Networks: Working with Participant-aided Sociograms. *Field Methods*, 19(2):116–144, 2007.
[30] Linked In. `www.linkedin.com`.
[31] T.S. Jayram, S. Kale, and E. Vee. Efficient aggregation algorithms for probabilistic data. In *SODA*, 2007.
[32] H. Kautz, B. Selman, and M. Shah. Referral web: combining social neworks and collaborative filtering. *Communications of the ACM*, 40(3), March 1997.
[33] A. C. Kerckhoff and K. W. Back. Sociometric patterns in hysterical contagion. *Sociometry*, 28(1):2–15, March 1965.

[34] D. Kirkpatrick. Facebook's plans to hookup the world. *Fortune Magazine, online edition*. `http://money.cnn.com/2007/05/24/technology/facebook.fortune/index.htm`, 2005.
[35] N. Kushmerick. Regression testing for wrapper maintenance. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, Orlando, Florida, July 1999. Menlo Park, CA: AAAI Press.
[36] A. L. Llyod and R. M. May. How viruses spread among computes and people. *Science*, 292(5520), May 2001.
[37] A. Machanavajjhala and J. Gehrke. On the efficiency of checking perfect privacy. In Stijn Vansummeren, editor, *PODS*, pages 163–172. ACM, 2006.
[38] D. McGuinness and A. Borgida. Explaining subsumption in description logics. In Chris Mellish, editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 816–821, San Francisco, 1995. Morgan Kaufmann.
[39] G. Miklau and D. Suciu. A formal analysis of information disclosure in data exchange. In *SIGMOD*, 2004.
[40] M. Morris. *Network Epidemiology: A Handbook for Survey Design and Data Collection*. Oxford University Press, 2004.
[41] M. Morris and M. Kretzschmar. Concurrent partnerships and the spread of HIV. *AIDS*, 11(5):641–648, 1997.
[42] MySpace. `www.myspace.com`.
[43] M. E. J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66(1):016128, Jul 2002.
[44] M. A. Nowak and R. May. *Virus dynamics: Mathematical principles of immunology and virology*. Oxford University Press, 2001.
[45] Visible Path. `www.visiblepath.com`.
[46] C. Ré, N. Dalvi, and D. Suciu. Query evaluation on probabilistic databases. *IEEE Data Engineering Bulletin*, 29(1):25–31, 2006.
[47] C. Ré, N. Dalvi, and D. Suciu. Efficient top-k query evaluation on probabilistic data. In *Proceedings of ICDE*, 2007.
[48] C. Ré and D. Suciu. Efficient evaluation of `HAVING` queries on probabilistic databases (full version). Technical report, University of Washington, Seattle, Washington, June 2007.
[49] C. Ré and D. Suciu. Materialized views in probabilsitic databases for information exchange and query optimization. In *VLDB '07 (to appear)*, 2007.
[50] B. Ryan and N. C. Gross. The diffusion of hybrid seed corn in two iowa communities. *Rural Sociology*, 8(1), 1943.
[51] A. D. Sarma, O. Benjelloun, A. Y. Halevy, and J. Widom. Working models for uncertain data. In Ling Liu, Andreas Reuter, Kyu-Young Whang, and Jianjun Zhang, editors, *ICDE*, page 7. IEEE Computer Society, 2006.
[52] P. Sen and A. Deshpande. Representing and querying correlated tuples in probabilistic databases. In *Proceedings of ICDE*, 2007.
[53] M. Soliman, I.F. Ilyas, and K. Chen-Chaun Chang. Top-k query processing in uncertain databases. In *Proceedings of ICDE*, 2007.
[54] T. W. Valente. *Network Models of the Diffusion of Innovations*. Hampton Press, 1995.
[55] T. W. Valente and R. L. Davis. Accelerating the diffusion of innovations using opinion leaders. *Annals of the American Academy of Political and Social Science*, 566, November 1999.
[56] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
[57] B. Wellman. Challenges in Collecting Personal Network Data: The Nature of Personal Network Analysis. *Field Methods*, 19(2):111–115, 2007.
[58] S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *Proceedings of KDD 2003*, pages 266–275, 2003.
[59] J. Widom. Trio: A system for integrated management of data, accuracy, and lineage. In *CIDR*, pages 262–276, 2005.
[60] M. Williams. The total information awareness project lives on. *Technology Review*, April 2006.
[61] W. Winkler. The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Bureau of the Census, 1999.