Exchangeable databases and their functional representation

James Robert Lloyd
Department of Engineering
University of Cambridge

Peter Orbanz
Department of Statistics
Columbia University

Zoubin Ghahramani
Department of Engineering
University of Cambridge

Daniel M. Roy
Department of Engineering
University of Cambridge

Abstract

We consider the task of statistical inference for data in the form of a relational database comprising multiple relations acting on heterogenous sets of objects. We define a notion of exchangeability for databases generalizing that of arrays, based on the idea that the objects over which the relations act are themselves exchangeable. When the data are encoded in the form of several multi-dimensional arrays this assumption corresponds to invariance to the simultaneous permutation of certain rows and columns across the multiple arrays. Recent work in Bayesian statistics has connected representation theorems due to Aldous, Hoover and Kallenberg to the modeling of individual exchangeable arrays. In particular, Hoff (2007), Roy and Teh (2009) and Lloyd et al. (2012) use these representational results to inspire statistical models of networks and other relational data. We extend this work by deriving corollaries of the representation theorems that are applicable to exchangeable databases and discuss the implications for modeling such data.

1 Introduction

Relational databases are an extremely common data structure so it is natural to want to perform statistical tasks with such data e.g., predicting unobserved data or identifying latent structure. In particular, network data is rarely encountered in isolation e.g., in a social network one will often have access to side information about each user. To perform a statistical analysis of such data we need to estimate parameters of a probabilistic model, but it is not immediately clear what an appropriate parameter space for such a model is. The choice of parameter space is important because it indicates the targets of statistical inference and shows where we can share statistical strength between different aspects of the data.

We demonstrate that the weak assumption of an appropriate form of exchangeability can provide a natural parameter space. This form of exchangeability is appropriate when the order of the objects underlying a relational database (e.g., users and movies in a database of ratings data) is arbitrary or unimportant. For example, the left hand side of figure 1 shows the same network but with differently labeled nodes. If the labeling is unimportant, then any probabilistic model of such data should assign them the same probability.

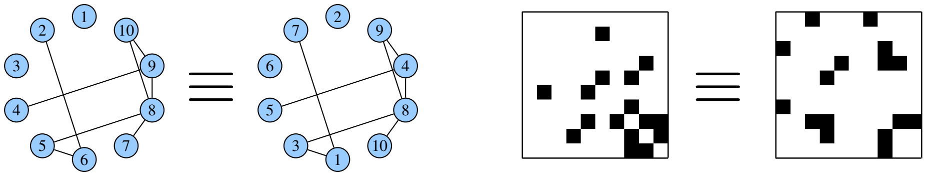

Figure 1: Left: Networks with equivalent structure but different node labels. Right: Corresponding adjacency matrix representations of these networks

Relational data are typically stored in arrays; the right hand side of figure 1 shows the corresponding adjacency matrix representations of the networks on the left. We demonstrate that exchangeability of the objects underlying a relational database can be expressed in terms of array exchangeability. Prior work on array exchangeability, both theoretical [e.g. 4, 5, 7–9, 15, 16, 18, 38] and applied [e.g. 13, 24, 30], has focused on single exchangaeble arrays. We show that the representation theorems for single arrays can be used to derive representations for collections of exchangeable arrays i.e., exchangeable databases.

2 Exchangeable databases

We abstractly define a database following the entity-relationship formalism [e.g. 36] where the values of attributes are the result of evaluating a function (relation) over a collection of entities / objects.

Definition 2.1 (types, signatures, relation)**.** Fix a finite set T of *types*. By a *signature* we mean a finite sequence $s \in T^n$ of types. By a *relation r of signature $s \in T^n$ with values in a space S* we mean a function from $\mathbb{N}^n$ to S.

We may encode a relation r with signature $s \in T^n$ as an array $X^r := (X^r_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^n}$ given by

$$X^r_{\boldsymbol{i}} = r(i_1, \dots, i_n), \qquad \text{for } \boldsymbol{i} = (i_1, \dots, i_n) \in \mathbb{N}^n. \tag{2.1}$$

Example 2.1. Let $T = \{\textit{users,movies}\}$. Then a relation r of signature $(\textit{users}, \textit{movies})$ taking values in $\{1, 2, 3, 4, 5\}$ might store movie ratings with rows corresponding to some enumeration of *users*, and columns corresponding to some enumeration of *movies*. A relation r' of signature $(\textit{users}, \textit{users})$ taking values in $\{0, 1\}$ might store the symmetric friendship relations in a social network.

Definition 2.2 (database)**.** By a *database* we mean a collection of R relations $r_1, \dots, r_R$ of signature $s_1, \dots, s_R$, respectively, taking values in $S_1, \dots, S_R$, respectively.

We may encode a database as a collection of arrays $(X^{r_j})_{j=1}^R$, where X^{r_j} encodes the relation r_j. For notational simplicity, we will often refer to the collection of arrays $(X^{r_j})_{j=1}^R$ as if it were the database itself.

Permuting the ordering of objects within a database results in a permutation of the indices of several of the arrays encoding its relations. We will now make this precise: For each type $t \in T$, let $p_t \in \mathbb{S}_\infty$ be a permutation of $\mathbb{N}$, i.e., $p_t : \mathbb{N} \to \mathbb{N}$ is one-to-one and onto. Write $p = (p_t; t \in T) \in \mathbb{S}_\infty^T$ for the collection of such permutations. Given a signature $s \in T^n$, define p^s to be the map from $\mathbb{N}^n$ to $\mathbb{N}^n$ such that

$$p^s(\boldsymbol{i}) := (p_{s_1}(i_1), \dots, p_{s_n}(i_n)), \qquad \text{for } \boldsymbol{i} \in \mathbb{N}^n. \tag{2.2}$$

In other words, p^s maps a sequence $i_1, \dots, i_n$ of indices (into the set of objects of type $s_1, \dots, s_n$, respectively) to the sequence where each index is permuted by the permutation corresponding to its type.

If X^r is the encoding of a relation r with signature $s \in T^n$, then the permuted relation $r \circ p$ is represented by the array $X^{r\circ p}$ given by

$$X^{r\circ p}_{\boldsymbol{i}} = X^r_{p^s(\boldsymbol{i})}, \qquad \text{for } \boldsymbol{i} \in \mathbb{N}^n. \tag{2.3}$$

Definition 2.3 (exchangeable database)**.** We say that a random database $(X^{r_j})_{j=1}^R$ is *exchangeable* when it has the same distribution as $(X^{r_j\circ p})_{j=1}^R$ for every $p \in \mathbb{S}_\infty^T$.

The following result characterizes the distribution of any exchangeable database to arbitrary accuracy. Let $(U_i^t)_{i \in \mathbb{N}, t \in T}$ be a collection of i.i.d. Uniform$[0,1]$ random variables.

Theorem 2.4 (functional representation for exchangeable databases)**.** *Let $(X^{r_j})_{j=1}^R$ be an exchangeable random database. Then there exists a sequence of random measurable functions $F^{j,1}, F^{j,2}, \dots$ for every $j = 1 \dots R$ such that the random databases $(X^{r_j,m})_{j=1}^R$ converge in distribution to $(X^{r_j})_{j=1 \dots R}$ as $m \to \infty$, where*

$$X_{\boldsymbol{i}}^{r_j,m} := F^{j,m}(U_{i_1}^{s_j(1)}, \dots, U_{i_n}^{s_j(n)}), \qquad \textit{for } \boldsymbol{i} \in \mathbb{N}^n. \tag{2.4}$$

The proof is based on a reduction to the representation theorem for π-exchangeable arrays presented in [18] but omitted for brevity. The theorem can be strengthened to state that the law of fixed subarrays are mutually absolutely continuous and the associated Radon-Nikodym derivatives converge uniformly to 1 as $n \to \infty$. An almost-sure representational result can also be given at the expense of heavy notation.

Because the notation is somewhat involved, we present a special case of this theorem applicable to e.g., modeling a network with side information for each node.

Corollary 2.5. *Consider an exchangeable database with one object type, one unary relationship, and one binary relationship; denote the binary relationship by the array $X = (X_{i,j})_{i,j \in \mathbb{N}}$ and the unary relationship with the sequence $C = (C_i)_{i \in \mathbb{N}}$. Then there exists a sequence of pairs of random measurable functions $(F^n, G^n)_{n \in \mathbb{N}}$ and a collection of i.i.d. Uniform$[0,1]$ random variables $(U_i)_{i \in \mathbb{N}}$ such that if we define the arrays $X^1, X^2, \dots$ and sequences $C^1, C^2, \dots$ by*

$$X_{i,j}^n := F^n(U_i, U_j), \qquad \textit{for } i, j, n \in \mathbb{N}, \tag{2.5}$$

$$C_{i,j}^n := G^n(U_i), \qquad \textit{for } i, n \in \mathbb{N}, \tag{2.6}$$

then (X^n, C^n) converges in distribution to (X, C) as $n \to \infty$.

Remark 2.6 (uniform distributions)**.** The uniform distributions in the theorem are canonical but the theorem still holds with any non-atomic probability measure on a Borel space e.g., Gaussian distributions.

2.1 Intepretation and examples

Theorem 2.4 states that the joint distribution of an exchangeable database can be arbitrarily well approximated by a collection of random measurable functions and uniform random variables. This functional form provides a set of parameters to be estimated that are naturally hierarchical. The functions $(F^{j,n})$ capture properties of entire relations whilst the (U_i^t) represent randomness associated with particular objects underlying the relational data.

2.1.1 Example : Exchangeable networks

Consider modeling a single binary relation which indicates whether or not two nodes in a network are connected or not. This data is typically represented in the form of an adjacency matrix (X_{ij}) where $X_{ij} = 1$ if and only if node i is connected to node j. Theorem 2.4 states that if the distribution of X is exchanegable then it can be arbitrarily well approximated by

$$(F(U_i, U_j)) \tag{2.7}$$

where F is a random measurable function and (U_i) are i.i.d. Uniform$[0,1]$ random variables. This special case has been used previously by [13, 24, 30] to inspire probabilistic models of networks of the form

$$\begin{aligned}
(U_i) &\sim_{\text{iid}} \text{e.g., Gaussian} && (2.8)\\
F &\sim \text{e.g., Gaussian process, bilinear function} \dots && (2.9)\\
W_{ij} &:= F(U_i, U_j) && (2.10)\\
X_{ij} | W_{ij} &\sim \text{Bernoulli}(\sigma(W_{ij})). && (2.11)
\end{aligned}$$

This is demonstrated pictorially in figure 2; in this case F can be interpreted as a blurred adjacency matrix.

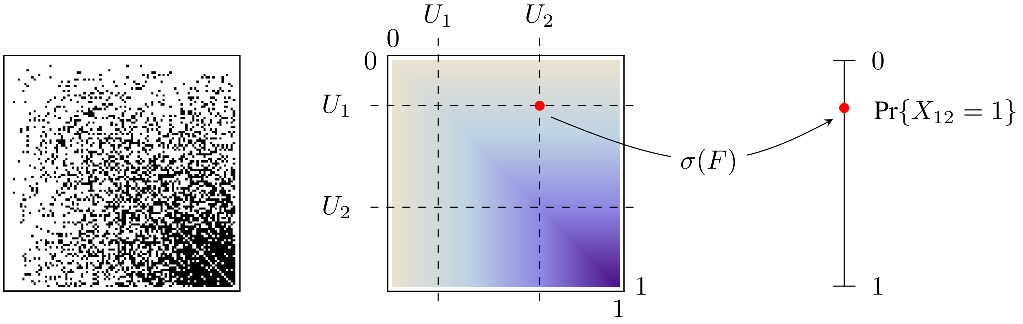

Figure 2: A pictorial representation of a model for network data inspired by the Aldous–Hover representation theorem. The left shows a random sample of a binary network (represented by an adjacency matrix) generated by a model of the form given by equation (2.11). This figure is adapted from [24] with permission.

2.1.2 Example : A simple database

Consider the simple database shown on the left hand side of figure 3. There are two objects, students and courses, and three relations, the unary relation 'age' acting on students, the binary relation 'friends' acting on pairs of students and the binary relation 'grade' that acts on students and courses. Sample data encoded in arrays is shown at the bottom of this figure.

Exchangeability of this database means that the entries of the leftmost table, the rows and columns of the second table and the rows of the third table can be arbitrarily permuted without changing the distribution of the database when viewed as a random variable. Similarly the columns of the rightmost table may be independently arbitrarily permuted.

The functional form resulting from the application of theorem 2.4 to this data structure is shown on the right hand side of figure 3. The two objects are represented by i.i.d. random variables, (U_i) for students, (V_i) for courses and the three relations are represented by three random functions $F(U_i), G(U_i, U_j)$ and $H(U_i, V_j)$ whose inputs are the random variables representing the objects the relations act upon.

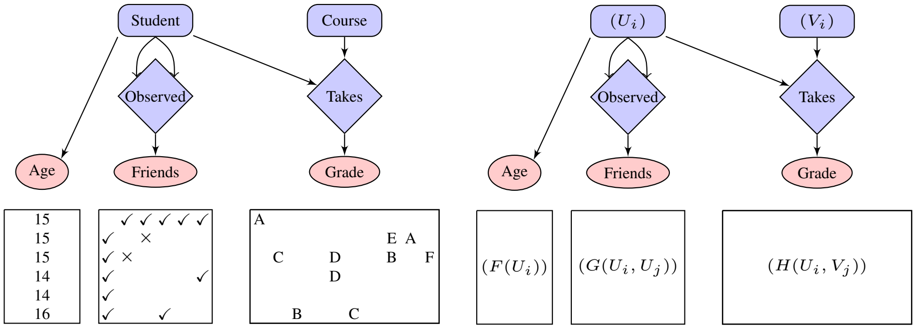

Figure 3: Left: A pictorial representation of a relational database. Right: The functional representation of the distribution of data of this form guaranteed to be an arbitrarily good approximation by theorem 2.4

3 A generic generative model

In analogy to the work of [13, 24, 30] on exchangeable arrays, theorem 2.4 naturally inspires a generic generative model of exchangeable databases. Each object in the database is associated with an i.i.d. sample, U_i^t, from some distribution $\mathcal{U}$ e.g., Uniform, Gaussian. For each relation $r_1, \ldots, r_R$ we independently sample a random function F^j from some distribution $\mathcal{F}^j$, e.g., Gaussian process, random (bi/tri/…)linear function. We denote the evaluation of these functions at the corresponding values of (U^t_i) by W^j. W^j can then be used to index into a family of observation distribution $L(\cdot)$, e.g., Gaussian distributions, to model the value of the relation r_j.

$$\begin{aligned}
(U^t_i) &\sim_{\text{iid}} \mathcal{U} && (3.1)\\
F^r &\sim \mathcal{F}^r && (3.2)\\
W^j_{\boldsymbol{i}} &:= F^j(U^{s_j(1)}_{i_1}, \cdots, U^{s_j(n)}_{i_n}) && (3.3)\\
X^{r_j}_{\boldsymbol{i}} \mid W &\sim L(W^j_{\boldsymbol{i}}) \qquad \text{independently across } j \text{ and } \boldsymbol{i}. && (3.4)
\end{aligned}$$

3.1 Prior work using models of this form

It was demonstrated in [24] that many models of single 2-arrays fit the form of the generic model presented above. In particular there are models that assume F is linear [e.g. 13, 25, 26, 31, 44], that F is Gaussian process distributed [e.g. 22, 24, 42] and other non-linear forms for F [e.g. 14, 30]. In addition to this there has been a line of work that uses increasingly more expressive forms of the distribution $\mathcal{U}$ [e.g. 20, 25, 26, 28, 29, 37, 40].

Many, but not all, of these models have been extended to model d-arrays. A summary of models using linear forms of F is given in [21]; non-linear models include [41].

For full databases, the literature is limited to clustering models [20] and models using linear forms for the F^r [e.g. 1–3, 6, 10–12, 17, 23, 27, 32–35, 39, 43].

4 Discussion

We have demonstrated how the concept of exchangeability can be applied to databases and used to derive a natural parameter space for statistical models of such data. Identifying a parameter space is the first step in any statistical analysis, allowing either frequentist estimation of the parameters or Bayesian prior specification. This concept is well established for exchanegable sequences where de Finetti's theorem [e.g. 19] applies. For exchangeable arrays, the relevant representation theorems were presented by Aldous and Hoover [4, 15] over 30 years ago but it is only recently that these results are being used to inspire Bayesian models [13, 24, 30] and frequentist estimation procedures [8, 18, 38]. We hope that this work will continue and be extended to the analysis of exchangeable databases.

References

[1] Acar, E., Kolda, T. G., and Dunlavy, D. M. (2011). All-at-once Optimization for Coupled Matrix and Tensor Factorizations. *arXiv preprint arXiv:1105.3422*.

[2] Acar, E., Plopper, G. E., and Yener, B. (2012). Coupled analysis of in vitro and histology tissue samples to quantify structure-function relationship. *PloS one*, **7**(3), e32227.

[3] Acar, E., Rasmussen, M. A., Savorani, F., Næ s, T., and Bro, R. (2013). Understanding Data Fusion Within the Framework of Coupled Matrix and Tensor Factorizations. *Chemometrics and Intelligent Laboratory Systems*.

[4] Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, **11**, 581—-598.

[5] Aldous, D. J. (2010). More uses of exchangeability: Representations of complex random structures. In *Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman*.

[6] Andersen, M. L. M., Rasmussen, M. A., Pörksen, S., Svensson, J., Vikre-Jø rgensen, J., Thomsen, J., Hertel, N. T., Johannesen, J., Pociot, F., Petersen, J. S., Hansen, L., Mortensen, H. B. l., and Nielsen, L. B. n. (2013). Complex Multi-Block Analysis Identifies New Immunologic and Genetic Disease Progression Patterns Associated with the Residual β-Cell Function 1 Year after Diagnosis of Type 1 Diabetes. *PloS one*, **8**(6), e64632.

[7] Austin, T. (2012). Exchangeable random arrays. *Technical report*.

[8] Choi, D. S. and Wolfe, P. J. (2013). Co-clustering separately exchangeable network data. *Annals of Statistics*.

[9] Diaconis, P. and Janson, S. (2007). Graph limits and exchangeable random graphs. *arXiv preprint arXiv:0712.2749*, pages 1–26.

[10] Ermis, B., Acar, E., and Cemgil, A. T. (2012). Link Prediction via Generalized Coupled Tensor Factorisation. *arXiv preprint arXiv:1208.6231*.

[11] Gallinari, P. and Upmc, L. I. P. (2011). Link Pattern Prediction with Tensor Decomposition in Multi-relational Networks. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*.

[12] Gao, S., Denoyer, L., and Gallinari, P. (2012). Probabilistic Latent Tensor Factorization Model for Link Pattern Prediction. In *3rd IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC)*.

[13] Hoff, P. D. (2007). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, pages 657–664.

[14] Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, **97**(460), 1090–1098.

[15] Hoover, D. N. (1979). Relations on Probability Spaces and Arrays of Random Variables. Technical report, Institute for Advanced Study, Princeton.

[16] Hoover, D. N. (1982). Row-column exchangeability and a generalized model for probability. In *Exchangeability in Probability and Statistics*, pages 281–291.

[17] Jimeng, Y.-r. L., Paul, S., Ravi, C., Hari, K., and Aisling, S. (2009). MetaFac : Community Discovery via Relational Hypergraph Factorization. pages 527–535.

[18] Kallenberg, O. (1999). Multivariate Sampling and the Estimation Problem for Exchangeable Arrays. *Journal of Theoretical Probability*, **12**(3), 859–883.

[19] Kallenberg, O. (2005). *Probabilistic symmetries and invariance principles*. Springer.

[20] Kemp, C., Tenenbaum, J., Griffiths, T., Yamada, T., and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21.

[21] Kolda, T. G. and Bader, B. W. (2009). Tensor Decompositions and Applications. *SIAM Review*, **51**(3), 455–500.

[22] Lawrence, N. D. and Urtasun, R. (2009). Non-linear matrix factorization with Gaussian processes. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1–8. ACM Press.

[23] Lippert, C., Weber, S. H., Huang, Y., Tresp, V., Schubert, M., and Kriegel, H.-p. (2008). Relation Prediction in Multi-Relational Domains using Matrix Factorization. (Siso).

[24] Lloyd, J. R., Orbanz, P., Roy, D. M., and Ghahramani, Z. (2012). Random function priors for exchangeable graphs and arrays. In *Advances in Neural Information Processing Systems (NIPS)*.

[25] Meeds, E., Ghahramani, Z., Neal, R. M., and Roweis, S. T. (2007). Modeling dyadic data with binary latent factors. In *Advances in Neural Information Processing Systems (NIPS)*. Citeseer.

[26] Miller, K. T., Griffiths, T. L., and Jordan, M. I. (2009). Nonparametric latent feature models for link prediction. *Advances in Neural Information Processing Systems (NIPS)*, pages 1276–1284.

[27] Nickel, M. (2011). A Three-Way Model for Collective Learning on Multi-Relational Data. In *Proceedings of the 28th international conference on machine learning (ICML-11)*.

[28] Nowicki, K. and Snijders, T. A. B. (2001). Estimation and Prediction for Stochastic Blockstructures. *Journal of the American Statistical Association*, **96**(455), 1077–1087.

[29] Palla, K., Knowles, D. A., and Ghahramani, Z. (2012). An Infinite Latent Attribute Model for Network Data. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[30] Roy, D. M. and Teh, Y. W. (2009). The Mondrian process. In *Advances in Neural Information Processing Systems*. Citeseer.

[31] Salakhutdinov, R. and Mnih, A. (2008). Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, pages 1–8.

[32] Shangguan, Q., Hu, L., Cao, J., and Xu, G. (2012). Book Recommendation Based on Joint Multi-relational Model. *2012 Second International Conference on Cloud and Green Computing*, pages 523–530.

[33] Singh, A. P. and Gordon, G. J. (2008a). A Unified View of Matrix Factorization Models. In *Machine Learning and Knowledge Discovery in Databases*, pages 358–373.

[34] Singh, A. P. and Gordon, G. J. (2008b). Relational Learning via Collective Matrix Factorization Categories and Subject Descriptors. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 650–658.

[35] Singh, A. P. and Gordon, G. J. (2012). A Bayesian Matrix Factorization Model for Relational Data. *arXiv preprint arXiv:1203.3517*.

[36] Ullman, J. and Widom, J. (2002). *A First Course in Database Systems*. Prentice Hall.

[37] Wang, Y. J. and Wong, G. Y. (1987). Stochastic Blockmodels for Directed Graphs. *Journal of the American Statistical Association*, **82**(397), 8–19.

[38] Wolfe, P. J. and Olhede, S. C. (2013). Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, pages 1–52.

[39] Xu, Q., Xiang, E. W., and Yang, Q. (2010). Protein-protein Interaction Prediction via Collective Matrix Factorization. In *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 62–67.

[40] Xu, Z., Tresp, V., and Yu, K. (2006). Infinite hidden relational models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

[41] Xu, Z., Yan, F., and Qi, Y. (2012). Infinite Tucker Decomposition: Nonparametric Bayesian Models for Multiway Data Analysis. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[42] Yan, F., Xu, Z., and Qi, Y. (2011). Sparse matrix-variate Gaussian process blockmodels for network modeling. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*.

[43] Yin, D., Guo, S., Chidlovskii, B., Davison, B. D., Archambeau, C., and Bouchard, G. (2013). Connecting Comments and Tags : Improved Modeling of Social Tagging Systems Categories and Subject Descriptors. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 547–556.

[44] Yu, K. and Chu, W. (2008). Gaussian process models for link analysis and transfer learning. In *Advances in Neural Information Processing Systems (NIPS)*.