Latent community discovery with network regularization for core actors clustering

XUN GuangXu[1], *YANG YuJiu*[2], *WANG LiangWei*[3], *LIU WenHuang*[4]
(1,2,4) Shenzhen Key Laboratory of Information Science and Technology,
Graduate School at Shenzhen, Tsinghua University, P.R.China
(3)Noah's Ark Laboratory, HUAWEI, Shenzhen, P.R.China
xgxo_atbash@yahoo.com.cn[1], {yang.yujiu, liuwh}@sz.tsinghua.edu.cn[2,4],
wangliangwei@huawei.com[3]

ABSTRACT
Community structure is a common attribute of many social networks, which would give us a better understanding of the networks. However, as the social networks grow lager and lager nowadays, the Pareto Principle becomes more and more notable which makes traditional community discovery algorithms no longer suitable for them. This principle explains the unbalanced existence of two different types of network actors. Specifically, the core actors usually occupy only a small proportion of the population, but have a large influence. In this paper, we propose a novel algorithm LCDN (Latent Community Discovery with Network Structure) for dividing the core actors. This is a hierarchical probabilistic model based on statistical topic model and regularized by network structure in data. We had experiments on three large networks which show that this new model performs much better than the traditional statistical models and network partitioning algorithms.

KEYWORDS: community discovery, statistical topic models, social networks, core actors, regularization.

Proceedings of COLING 2012: Posters, pages 1351–1360,
COLING 2012, Mumbai, December 2012.

1 Introduction

Social network has been studied for a long time in both empirical ways and theoretical ways. A common attribute of many networks is community structure. Discovering this inherent attribute can lead us to a deeper understanding of the networks (Scott, 1988). The study of community structure in networks is mainly related to the graph partitioning of graph theory and statistical model. Most of the graph partitioning and statistical model algorithms have been proved effective. However, a new problem which is known as the Pareto principle arose, especially in large networks. For many events, roughly 80% of the effects come from 20% of the causes (Wikipedia, 2001). This also fits many large social networks. In these networks, there exist two different kinds of actors with disparate social behaviors and social influence. The core actors get a small proportion of population but make a big proportion of social influence. See figure 1 for an example. Core actors get more attention than the ordinary ones.

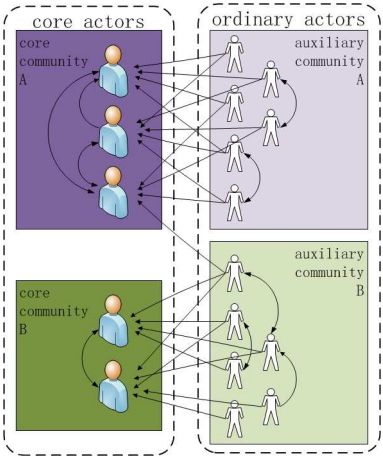

Figure 1: Two different kinds of actors

Because of the Pareto principle, the existed community discovery algorithms do not perform very well about core actors. In order to address this problem, we design a regularized model LCDN (Latent Community Discovery with Network Structure). The rest of this paper is organized as follows. We present related work in section 2. In section 3, we define the problem of community discovery on core network. And In section 4, we propose the novel algorithm LCDN. Finally in section 5 we discuss the experiments and evaluation.

2 Related work

Community discovery is a problem that arise in, for example, the Social Network Service (SNS). It is mainly related to the graph partitioning of graph theory and statistical model. For the graph partitioning of graph theory, its solutions fall into two main classes, agglomerative and divisive, depending on whether the procedure is to add or to remove the edges in the network to form communities, such as the k-means algorithm (Hartigan and Wong, 1979) and the

Girvan-Newman algorithm (Newman and Girvan, 2004). Statistical model is another way to discover community structure (Zhang et al., 2007b). Statistical model, especially topic model, has been applied to many domains such as information retrieval successfully (Chang and Blei, 2009). Compared with the graph partitioning of graph theory, statistical model for community detection introduces probability, which means one actor in the network could belong to more than one community and the boundaries between different communities could be blurry (Zhang et al., 2007a). That makes it more explainable (Chang et al., 2009).

3 Problem formulation

We assume that the network to be analyzed is influenced by Pareto Efficiency. These networks that we are going to handle conspicuously separated the actors into two groups, namely the core actors and the ordinary actors. Now we introduce the related definitions of LCDN:
Definition 1 (core actor): core actors are more influential. For example, in scientific coauthors, they publish most of the papers, and in a SNS, they get the highest in-degrees.
Definition 2 (association document): an association document d in a network n is a sequence of core actors $a_1, a_2 ... a_{|d|}$ that are associated with the current actor, where a_i is an element from a fixed core actor map. And c(a,d) means the occurrences of actor a in d.
Definition 3 (core network): a network is a graph G=<V, E>, where V is a set of vertices and E is a set of edges. A vertex $u \in V$ represents a single actor of the network associated with its association document. An edge <u, v> is a connection between vertices u and v. It can be either directed or undirected. A core network is an extraction of the whole network, a sub network that its vertices only consist of the core actors.
Definition 4 (latent community): a latent community in our model corresponds to a topic in the topic model. We represent it with z, then we have $\sum_z P(z|d) = 1$ and $\sum_a P(a|z) = 1$. And we assume that there are k latent communities in this network.
Definition 5 (community map): for each core actor in the association document, its community map is a weighting function $f(z, a)$ that shows the probabilistic relevance between core actor and latent community. For example, we may define $f(z, a) = P(z|a)$. And we expect that the adjacent actors have similar community maps.

4 Latent Community Discovery with Network Structure

4.1 Probabilistic latent semantic analysis

First of all, we introduce the PLSA (Probabilistic Latent Semantic Analysis) topic model which our statistical part is based on. The PLSA model assumes that there are k topics in the corpora, where k is a fixed parameter, and every document in the corpora corresponds to one distribution of topics. This is a hierarchical model. We can describe its generative process as:

- Select a document d with probability $P(d)$.
- Pick a latent topic z with probability $P(z|d)$.
- Generate a word w with probability $P(w|z)$.

So we obtain an observed pair $P(d, w_n) = P(d) \sum_z P(z|d)P(w_n|z)$ (Hofmann, 1999). When this statistical model is used to do community detecting, documents are replaced by association documents, and words in a document correspond to actors in an association document. So, the log likelihood of a network n to be generated with PLSA model is given by:

$$L = \sum_{d=1}^{M} \sum_{a=1}^{N_d} C(a, d) * \log \sum_{j=1}^{T} P(z_j|d) * P(a|z_j) \quad (1)$$

4.2 The probabilistic community discovery framework

By regularizing the statistical model with network structure, we propose a framework that takes both the statistical model and the network structure into consideration. As we expect that the adjacent actors have similar community maps, the criterion function is succinct and natural:

$$O(N,G) = x * (-L(N)) + (1-x) * R(N,G) \tag{2}$$

Where $L(N)$ is the log likelihood of an association corpora with statistical model, and $R(N,G)$ is the regularizer defined on the network structure. This criterion function is very general, where $L(N)$ can be the log likelihood of any statistical model and $R(N,G)$ can be any regularizer to make the community map smoother. In this paper, we choose PLSA as the statistical model, and define the regularizer $R(N,G)$ as (Zhu et al., 2003):

$$R(N,G) = \frac{1}{2} \sum_{(a_1,a_2)\in E} \sum_{z} (P(z|a_1) - P(z|a_2))^2 \tag{3}$$

By maximizing $L(N)$, we will get the $P(z|d)$ and $P(a|z)$ that fit our association document the best, and by minimizing $R(N,G)$, we will get the $P(z|a)$ that smooth our network structure the most. The parameter x will be set between 0 and 1 to control the balance between statistical log likelihood and smooth regularizer. Then we have the following optimization problem:

$$\begin{aligned} O(N,G) = x * (-\sum_{d=1}^{M} \sum_{a=1}^{N_d} C(a,d) * \log \sum_{j=1}^{T} P(z_j|d) * P(a|z_j)) \\ + (1-x) * \frac{1}{2} \sum_{(a_1,a_2)\in E} \sum_{z} (P(z|a_1) - P(z|a_2))^2 \end{aligned} \tag{4}$$

4.3 Parameter inference

When $x = 0$, the criterion function turns into a standard log likelihood of the PLSA model. The way to infer and estimate parameters for PLSA is the EM (Expectation Maximization) algorithm (Dempster et al., 1977), so we can find a local maximum of $L(N)$ in this iterative way. In the PLSA model, the E-step boils down to computing the probability of latent variables:

$$P(z|a,d) = \frac{P(z|d) * P(a|z)}{\sum_{j'=1}^{T} P(z_{j'}|d) * P(a|z_{j'})} \tag{5}$$

Take the latent variables into consideration, and then we have its complete likelihood:

$$Q(N) = \sum_{d=1}^{M} \sum_{a=1}^{N_d} C(a,d) * \sum_{z} P(z|a,d) \log(P(z|d)P(a|z)) \tag{6}$$

By maximizing the complete likelihood in the-M step, we obtain the following updated equations:

$$P(a|z) = \frac{\sum_d C(a,d) * P(z|a,d)}{\sum_{d,a} C(a,d) * P(z|a,d)} \qquad P(z|d) = \frac{\sum_a C(a,d) * P(z|a,d)}{\sum_{a,z} C(a,d) * P(z|a,d)} \tag{7}$$

When $x \neq 0$, it becomes more complicated. Then we consider the constraints:

$$\sum_z P(z|d) - 1 = 0, \qquad \sum_a P(a|z) - 1 = 0, \qquad \sum_z P(z|a) - 1 = 0$$

Add Lagrange multipliers corresponding to the constraints, we obtain the following complete likelihood with network structure information:

$$\begin{aligned} Q(N,G) = x * (-\sum_{d=1}^{M}\sum_{a=1}^{N_d} C(a,d) * \sum_z P(z|a,d)\log(P(z|d)P(a|z))) \\ + \sum_d \alpha_d(\sum_z P(z|d) - 1) + \sum_z \alpha_z(\sum_a P(a|z) - 1) + \sum_a \alpha_a(\sum_z P(z|a) - 1) \\ + (1-x) * \frac{1}{2} \sum_{(a_1,a_2)\in E} \sum_z (P(z|a_1) - P(z|a_2))^2 \end{aligned} \tag{8}$$

Where α_d,α_z and α_a are all Lagrange multipliers. Continue our EM process to seek the local minimum of $Q(N,G)$. It is easy to see that the latent variables do not change from equation 6 to equation 9 compared with PLSA model, so the E-step of LCDN is still the same as equation 5. The introduction of network structural $R(N,G)$ do not affect the $P(z|d)$, so the estimation of $P(z|d)$ remains as equation 8. However, $P(a|z)$ and $P(z|a)$ are no longer calculable directly by minimizing $Q(N,G)$. The Newton-Raphson method is a good way to solve this kind of problem. Suppose x_n is the variable to be updated by the Newton-Raphson method at n-th iteration, corresponding to the unknown parameters $P(a|z)$ and $P(z|a)$ in our model. Specifically applied to our task:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - HQ(x_n;N,G)^{-1} \bigtriangledown Q(x_n;N,G) \tag{9}$$

Where $\bigtriangledown Q(x_n;N,G)$ is the gradient of $Q(x_n;N,G)$ and $HQ(x_n;N,G)$ is the Hessian matrix of $Q(x_n;N,G)$.

4.4 An efficient algorithm

In fact, we could achieve the expected effect by ensuring $Q_{n+1}(N,G) < Q_n(N,G)$ at every M-step. So we can optimize the statistical complete likelihood part and network structural regularizer of the objective function separately. In each M-step, we could maximize the complete likelihood by equation 7 and equation 8 as before, but this does not necessarily mean $Q_{n+1}(N,G) < Q_n(N,G)$, so we have to continue optimizing the structural part $R(N,G)$. Obviously, random walk on the network is a simple and effective choice to minimize $R(N,G)$. Thus for $P(z|a)$:

$$P_{n+1}(z|a) = x * P_n(z|a) + (1-x) * \frac{\sum_{(a,a')\in E} C(a,a') * P_n(z|a')}{\sum_{(a,a')\in E} C(a,a')} \tag{10}$$

It is easy to see that $\sum_z P_{n+1}(z|a) = 1$ and $P_{n+1}(z|a) \geq 0$ always hold in equation 11. Here, x is the random walk parameter. Every iteration of random walk makes the network smoother (Jamali and Ester, 2009). Note that, the random walk process is based on $P(z|a)$ of each actor in the network, so we have to use the Bayes' theorem to obtain $P(a|z)$ for next EM algorithm iteration:

$$P(a|z) = \frac{P(z|a) * P(a)}{P(z)} = \frac{P(z|a) * P(a)}{\sum_a P(z|a) * P(a)} \tag{11}$$

5 Experiments and evaluation

In this section, we experiment on three large networks. Since the LCDN is unsupervised, pairwise comparison is a good choice to measure our experiment results (Menestrina et al., 2010). So we get the pairwise $precision$, $recall$ and $F1$:

$$precision(E,G) = \frac{|pair_E \bigcap pair_G|}{|pair_E|} \qquad recall(E,G) = \frac{|pair_E \bigcap pair_G|}{|pair_G|} \tag{12}$$

$$F1(E,G) = \frac{|2 * precision * recall|}{|precision + recall|} \tag{13}$$

Besides, in order to obtain a more comprehensive evaluation of the model, we add two comparison indexes: time cost to measure efficiency and community size to measure partitioning balance. We pick two statistical models (PLSA and Ball-Karrer-Newman models) and two graph partitioning models (k-means and Newman-Girvan models) as the comparative algorithms.

5.1 DBLP co-authorship network

In DBLP co-authorship network, the actors represent authors, and the edges represent the collaboration relation between authors. This benchmark co-authorship network contains the co-authorship of more than 50000 papers published at 27 computer science conferences from 2008 to 2010. And the whole network gets 59073 actors and 151399 edges. These conferences can be mainly divided into five research groups (Wang et al., 2011):

1. AI: artificial intelligence, including IJCAI, AAAI, ICML, UAI, NIPS, UMAP and AAMAS.
2. DB: database, including EDBT, ICDT, ICDE, PODS, SIGMOD and VLDB.
3. DP: distributed and parallel computing, including ICCP, IPDPS, PACT, PPoPP and Euro-Par.
4. GV: graphics, vision and HCI, including I3DG, ICCV, CVPR and SIGGRAPH.
5. NC: networks communications and performance, including MOBICOM, INFOCOM, SIGMETRICS, PERFORMANCE and SIGCOMM.

Intuitively, we believe that the program committee members are academically active in their respective areas, so that these program committee members (1241 actors, including 406 AI members, 282 DB members, 323 NC members, 79 GV members and 188 DP members) constitute the core actor group. In PLSA and LCDN model, the input weight in the association document is the number of collaboration times between current actor and target actor. And NG is short for Newman-Girvan algorithm and BKN is short for Ball-Karrer-Newman algorithm. The core community discovery result comparison of these algorithms is given in table 1:

	Pairwise precision	Pairwise recall	Pairwise F1	Time cost(s)	Community size				
					C1	C2	C3	C4	C5
PLSA	0.276	0.238	0.265	**19**	243	199	244	256	299
k-means	0.257	**0.978**	0.406	569	2	19	1218	1	1
NG	Unavailable since this network is not fully interconnected								
BKN	0.156	0.306	0.206	799	852	100	80	126	83
LCDN	**0.456**	0.434	**0.445**	114	**136**	**370**	**108**	**251**	**376**

Table 1: Algorithm performance comparison on the DBLP co-authorship network

And then we move on to the fully interconnected DBLP co-authorship network. This is the biggest sub network extracted from the DBLP co-authorship network, whose actors are fully interconnected. So the Newman-Girvan algorithm would work on it. This extraction contains 42956 actors and 130132 edges. And the number of core actors is reduced to 1222, including 402 AI members, 281 DB members, 319 NC members, 79 GV members and 178 DB members. The core community discovery result comparison of these algorithms is given in table 2:

	Pairwise precision	Pairwise recall	Pairwise F1	Time cost(s)	Community size				
					C1	C2	C3	C4	C5
PLSA	0.309	0.245	0.273	**18**	221	252	242	213	294
k-means	0.258	0.979	0.409	432	1199	19	1	1	2
NG	0.253	**0.988**	0.403	8625	1217	1	1	1	2
BKN	0.287	0.480	0.359	755	764	102	104	125	127
LCDN	**0.501**	0.465	**0.483**	267	**394**	**298**	**171**	**92**	**267**

Table 2: Comparison on the fully interconnected DBLP co-authorship network

We can see from table 1 and 2 that LCDN achieves an 8% $\sim$ 20% improvement in terms of pairwise $F1$ value over the other comparative algorithms. And the k-means algorithm and Newman-Girvan algorithm are confronted with the problem of unbalanced partitioning.

5.2 WEIBO network

Compared with the DBLP co-authorship network, WEIBO network is a much larger one with 164524 actors and 14444794 edges. WEIBO is a directed graph whose actors are fully interconnected. And there are two kinds of edges in it, strong and weak. This will influence the actor weight in association documents. In this paper, we set the weight to 1 for the weak ones and to 11 for the strong ones. Since the actors of WEIBO network get far more neighbors than the ones of DBLP co-authorship network, the random walk parameter is set smaller to keep the regularization balance of LCDN model. And in fact, to achieve a better performance, the denser the network, the smaller value of random walk parameter x we should set. In this paper, we set the random walk parameter of DBLP co-authorship network to $x = 0.9$, and set the random walk parameter of WEIBO network to $x = 0.1$.

Intuitively, we believe that in a SNS the more followers mean the more influence, so we pick actors whose in-degree is greater than 2000 to constitute the core actor group. These actors can be mainly divided into 5 social groups according to their tags and verification information:

1. Entertainment and sports, including 244 members.
2. Grass roots and leisure, including 333 members.
3. Finance and technology, including 297 members.
4. Culture and religion, including 185 members.
5. Newspapers and media, including 163 members.

The core community discovery result comparison of these algorithms is given in table 3:

We can see that LCDN still gets a much better performance than the other algorithms. For the k-means algorithm and Newman-Girvan algorithm, the problem of unbalanced partitioning remains. Besides, compared with the other algorithms, the time cost of Newman-Girvan algorithm is really unacceptable. Empirically, the Newman-Girvan algorithm should not partition a network so unbalanced as a divisive method. So we traced the actor Li Kaifu, who had the

	Pairwise precision	Pairwise recall	Pairwise F1	Time cost(s)	Community size				
					C1	C2	C3	C4	C5
PLSA	0.627	0.567	0.596	1098	176	200	235	299	271
k-means	0.227	0.715	0.345	**558**	77	108	2	992	2
NG	0.201	**0.873**	0.326	85084	11	1124	21	10	15
BKN	0.528	0.478	0.502	4164	184	270	227	304	196
LCDN	**0.682**	0.611	**0.645**	2067	**282**	**185**	**221**	**277**	**216**

Table 3: Algorithm performance comparison on the WEIBO network

highest in-degree in WEIBO network but was assigned to a small community. Actually, we found that, the community that Li Kaifu was assigned to was not really small because it also contained thousands of ordinary actors. As Li Kaifu and other famous actors have numerous followers, this will surely lead to an very high betweenness score of the edges between these core actors. Thus edges between core actors tend to be cut with a very high priority, and it is indeed so according to our observation of Li.

Conclusion and perspectives

A structure of communities with Pareto Principle exists in many real-world social networks. Every individual does not get the same social influence. Naturally, we pay more attention to the core actors, since they are the kernel of a network. In this paper, we define the problem of community detection among the core actors in large social networks. Taking both the statistical model and the network structure into consideration, we propose a probabilistic community discovery framework LCDN. The experimental results show that LCDN model performs much better than the other algorithms.

For future work, we would like to try to make this framework multifunctional, for example to collaborative filtering, and develop this framework into a fuller Bayesian model. Since we can obtain the association document parameters $P(z|d)$ which could properly represent the interest of current actor. So we can do collaborative recommendation based on either the community map or association document parameter $P(z|d)$ (Su and Khoshgoftaar, 2009). The PLSA model gets limitations that there is no constraint on the association document parameters $P(z|d)$. This leads to overfitting: the number of association document parameters grows linearly with the data size (Mei et al., 2008). The LDA (Latent Dirichlet Allocation) model is a good choice to alleviate this problem (Blei et al., 2003; Griffiths, 2002). Moreover, the LDA model is more general. It gets plenty of variations which pay different emphases so that it is applicable to many different situations (Ramage et al., 2009; Blei and Lafferty, 2006; Blei and McAuliffe, 2007; Blei and Lafferty, 2005).

Acknowledgments

The work was supported by Guangdong Natural Science Foundation (No.9451805702004046) and the cooperation project in industry, education and research of Guangdong province and Ministry of Education of P.R. China (No.2010B090400527). In addition, we thank the anonymous reviewers for their careful read and valuable comments.

References

Blei, D. M. and Lafferty, J. D. (2005). Correlated topic models. In *NIPS*.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In Cohen, W. W. and Moore, A., editors, *ICML*, volume 148 of *ACM International Conference Proceeding Series*, pages 113–120. ACM.

Blei, D. M. and McAuliffe, J. D. (2007). Supervised topic models. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *NIPS*. Curran Associates, Inc.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Chang, J. and Blei, D. M. (2009). Relational topic models for document networks. *Journal of Machine Learning Research - Proceedings Track*, 5:81–88.

Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A., editors, *NIPS*, pages 288–296. Curran Associates, Inc.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

Griffiths, T. (2002). Gibbs sampling in the generative model of latent dirichlet allocation.

Hartigan, J. A. and Wong, M. A. (1979). A k-means clustering algorithm. *Journal of the Royal Statistical Society*, 28(1):100–108.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57. ACM.

Jamali, M. and Ester, M. (2009). *TrustWalker*: a random walk model for combining trust-based and item-based recommendation. In IV, J. F. E., Fogelman-Soulié, F., Flach, P. A., and Zaki, M. J., editors, *KDD*, pages 397–406. ACM.

Mei, Q., Cai, D., Zhang, D., and Zhai, C. (2008). Topic modeling with network regularization. In Huai, J., Chen, R., Hon, H.-W., Liu, Y., Ma, W.-Y., Tomkins, A., and Zhang, X., editors, *WWW*, pages 101–110. ACM.

Menestrina, D., Whang, S., and Garcia-Molina, H. (2010). Evaluating entity resolution results. *PVLDB*, 3(1):208–219.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, pages 248–256. ACL.

Scott, J. (1988). Social network analysis. *Sociology*, 22(1):109–127.

Steyvers, M. and Griffiths, T. (2007). *Probabilistic Topic Models*. Handbook of Latent Semantic Analysis.

Su, X. and Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Adv. Artificial Intellegence*, 2009.

Wang, L., Lou, T., Tang, J., and Hopcroft, J. E. (2011). Detecting community kernels in large social networks. In Cook, D. J., Pei, J., Wang, W., Zaïane, O. R., and Wu, X., editors, *ICDM*, pages 784–793. IEEE.

Wikipedia (2001). Pareto principle. `http://en.wikipedia.org/wiki/Pareto_principle`.

Zhang, H., Giles, C. L., Foley, H. C., and Yen, J. (2007a). Probabilistic community discovery using hierarchical latent gaussian mixture model. In *AAAI*, pages 663–668. AAAI Press.

Zhang, H., Qiu, B., Giles, C. L., Foley, H. C., and Yen, J. (2007b). An lda-based community structure discovery approach for large-scale social networks. In *ISI*, pages 200–207. IEEE.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In Fawcett, T. and Mishra, N., editors, *ICML*, pages 912–919. AAAI Press.