(Un)detectable Cluster Structure in Sparse Networks

Jörg Reichardt[1] and Michele Leone[2]
[1]Institute for Theoretical Physics, University of Würzburg, 97074 Würzburg, Germany
[2]ISI Foundation, Viale S. Severo 65,I-10133 Torino, Italy
(Received 9 November 2007; published 13 August 2008)

Can a cluster structure in a sparse relational data set, i.e., a network, be detected at all by unsupervised clustering techniques? We answer this question by means of statistical mechanics making our analysis independent of any particular algorithm used for clustering. We find a sharp transition from a phase in which the cluster structure is not detectable at all to a phase in which it can be detected with high accuracy. We calculate the transition point and the shape of the transition, i.e., the theoretically achievable accuracy, analytically. This illuminates theoretical limitations of data mining in networks and allows for an understanding and evaluation of the performance of a variety of algorithms.

DOI: 10.1103/PhysRevLett.101.078701 PACS numbers: 89.75.Hc, 05.50.+q, 89.65.−s

Clustering is of fundamental importance in exploratory data analysis. We ask whether clusters in sparse networks can be recovered at all and what is the maximum achievable accuracy for any algorithm [1]. For example, consider clustering the pages of the world wide web thematically. Pages on common subjects will be linked more densely than pages on different subjects. Under what circumstances is it possible to infer these clusters corresponding to different topics from the link structure alone?

We study an ensemble of clustered networks. All nodes $i \in \{1, \ldots, N\}$ are assigned into one of q designed or ''planted'' clusters of given size n_s with $\sum_s^q n_s = N$ and carry a hidden variable $s_i \in \{1, \ldots, q\}$ indicating this cluster. The degree distribution $p^s(k)$ may differ between clusters, but the average connectivity per node $\langle k\rangle^s = \sum_{k=1}^{\infty} k p^s(k)$ is finite for all s in accordance with real world networks [2,3]. Links are distributed randomly yet obeying $p^s(k)$ and a matrix of conditional probabilities $p(r|s)$ parametrizing the cluster structure. Here, $p(r|s)$ denotes the conditional probability that given a link with one end in designed cluster s its other end belongs to designed cluster r. We only consider parameters which obey $p(r|r) > \langle k\rangle^r n_r/2M \;\forall\; r$ and $p(s|r) < \langle k\rangle^s n_s/2M \;\forall\; s \neq r$, i.e., we want more links within clusters than expected from a purely random assignment of nodes into clusters. Here, M is the total number of links in the network.

Clustering means to infer labels for nodes $\sigma_i \in \{1, \ldots, q\}$. The accuracy of recovering the hidden cluster labels is measured by $A = \sum_i \delta_{s_i,\sigma_i}/N$. Given are *only* the links in the network, the number of clusters, and their respective sizes. With unknown $p(r|s)$, any algorithm must partition the nodes of the network into q groups of given size minimizing the number of links between different groups, i.e., search for maximally separated clusters. This corresponds to a minimum cut partitioning problem. For $p(r|r) = 1 \;\forall\; r$ the network consists of q disconnected components and inference is trivial. For $p(s|r) = \langle k\rangle^s n_s/2M \;\forall\; r, s$ cluster structure is absent and inference of clusters is impossible.

In this Letter we study the transition from the impossible to the trivial case as a function of $p(r|s)$. For simplicity, in the following we will restrict the analysis to the case of q equal sized clusters which all have the same degree distribution $p^s(k) = p(k) \;\forall\; s$, such that $p(r|r) = p_{\text{in}} \;\forall\; r$ and $p(s|r) = (1 - p_{\text{in}})/(q - 1) \;\forall\; s \neq r$. We will find that feasible inference is only possible if p_{in} is larger than a critical value p_{in}^c. Moreover, this p_{in}^c depends crucially on the degree distribution of the network. We will calculate this dependence analytically. Finally, we will calculate the maximum achievable accuracy that any algorithm can reach on the ensemble of networks we describe.

A minimum cut partition is a ground state of the following ferromagnetic Potts Hamiltonian:

$$\mathcal{H} = -\sum_{i<j} J_{ij}\delta_{\sigma_i,\sigma_j} \tag{1}$$

under the constraint of the given cluster sizes. The constraint of an equipartition is enforced by a symmetry condition on the order parameters of the system (see below). The couplings J_{ij} are the entries of the $\{0, 1\}$ adjacency matrix of the graph. In the ground state, the σ_i are the inferred cluster labels. From the energy per node E in the ground state, one may calculate a number of commonly used quality functions for clustering such as the modularity $\mathcal{Q} = -2E/\langle k\rangle - 1/q$ of this partition [4].

The preassigned partition is guaranteed to be a minimum cut only for $p_{\text{in}} = 1$. Generally, the energy $E^p = -p_{\text{in}}\langle k\rangle/2$ of the preassigned partition ($\sigma_i = s_i$) is larger than the ground state energy of (1). The question we ask is whether the ground state configuration of (1) is influenced by the clustered topology of our network and thus overlaps with the hidden cluster labels. To answer, we need to calculate the local field distribution at each site in the ground state of (1) dependent on the preassigned cluster index. In the limit of $p_{\text{in}} = 1/q$, i.e., without a cluster structure, this problem has been studied extensively for Poissonian degree distributions or Bethe lattices with a fixed valence [5–9].

To calculate the field distributions we employ the Bethe-Peirls approach (or cavity method or belief propagation), directly at zero temperature [10]. The distribution of local fields $P^s(\mathbf{h})$ is calculated from an integral over a distribution of cavity biases or "messages" $Q^s(\mathbf{u})$ which are calculated self-consistently. The superscript $s \in \{1, \ldots, q\}$ denotes a possible dependence of these distributions on the cluster index of the preassigned cluster structure. An easy to follow derivation of these equations can be found in Refs. [11,12]. The equations read

$$Q^s(\mathbf{u}) = \sum_{d=0}^{\infty} q^s(d) \int \prod_{i=1}^{d} [d^q \mathbf{u}_i Q^s_{\text{in}}(\mathbf{u}_i)] \delta\Big(\mathbf{u} - \hat{\mathbf{u}}\Big(\sum_{i=1}^{d} \mathbf{u}_i\Big)\Big),$$
$$P^s(\mathbf{h}) = \sum_{k=0}^{\infty} p^s(k) \int \prod_{i=1}^{k} [d^q \mathbf{u}_i Q^s_{\text{in}}(\mathbf{u}_i)] \delta\Big(\mathbf{h} - \sum_{i=1}^{k} \mathbf{u}_i\Big). \quad (2)$$

Here, $q^s(d) = (d+1)p^s(d+1)/\langle k \rangle^s$ is the distribution of the excess degree per node. These equations are solved by iteration or population dynamics, often called "message passing." The topology of the clustered network enters via the distribution of "incoming" messages $Q^s_{\text{in}}(\mathbf{u}) = \sum_r^q p(r|s) Q^r(\mathbf{u})$. Equations (2) are general as the particular form of the Hamiltonian enters only via the two functions $v(\mathbf{h})$ and $\hat{\mathbf{u}}(\mathbf{h})$. For the Hamiltonian (1) the function $\hat{\mathbf{u}}$ is defined via

$$v(\mathbf{h}) = \max(h^1, \ldots, h^q), \quad (3)$$

$$\hat{u}^s(\mathbf{h}) = \max(h^1, h^s + 1, \ldots, h^q) - v(\mathbf{h}). \quad (4)$$

This means that $\hat{\mathbf{u}}$ picks the maximum components in $\mathbf{h}$ and sets all corresponding components in $\mathbf{u}$ to one and the rest to zero. Because of possible degeneracy in the components of $\mathbf{h}$, the vector $\mathbf{u} = \hat{\mathbf{u}}(\mathbf{h})$ may have more than one nonzero entry and is never completely zero. This observation is fundamental for all further developments. The components of $\mathbf{h}$ take only integer values, because we only have integer couplings J_{ij}.

Under the assumption of replica symmetry, the above approach is exact on an infinitely large graph. The solutions are hence approximations for the field distributions in a finite graph with the same degree distribution.

There are $2^q - 1$ possible messages $\mathbf{u}$. The probabilities of sending them may depend on the planted cluster from which they are sent, hence there are $q(2^q - 1)$ different probabilities $Q^s(\mathbf{u})$ to determine. We are only interested in distributions that allow to fulfill the constraint of an equipartition and that are symmetric under permutation of the indices as is our planted cluster structure. These conditions reduce the number of different probabilities $Q^s(\mathbf{u})$ to only $2q - 1$ order parameters η_{cw}:

$$Q^s(\mathbf{u}) = \eta_{\text{cw}}, \quad \text{where } c = u^s \quad \text{and} \quad w = \|\mathbf{u}\|^2 - c. \quad (5)$$

Here, u^s denotes the sth component of the message vector $\mathbf{u}$ under consideration. Without loss of generality, we have thus introduced a preferred direction for each planted cluster. The probability $Q^s(\mathbf{u})$ that a node from planted cluster s sends a message $\mathbf{u}$ depends only on whether or not $\mathbf{u}$ has an entry of one in the "correct" component s ($c = 1$) and on how many "wrong" components w in $\mathbf{u}$ carry an entry of one ($w \in \{1 - c, \ldots, q - 1\}$). For $p_{\text{in}} \to 1$ we must have $\eta_{10} \to 1$, i.e., only correct messages are sent. For $p_{\text{in}} \to 1/q$ we must have $\eta_{1,\alpha-1} = \eta_{0,\alpha} = \eta_\tau$, i.e., the probability of a message depends only on the number $\tau = w + c$ of nonzero entries in it. These new order parameters describing $Q^s(\mathbf{u})$ obey

$$\sum_{c=0}^{1} \sum_{w=1-c}^{q-1} \binom{q-1}{w} \eta_{\text{cw}} = 1. \quad (6)$$

Let us now turn to the case of two clusters. Then, we only have three possible messages $\mathbf{u} \in \{(1,0), (0,1), (1,1)\}$ and three order parameters η_{cw}. The equation for $Q^s(\mathbf{u})$ can then be written as a set of polynomial equations for the η_{cw} in a simple way:

$$\eta_{11} = \sum_{n0=0}^{\infty} \sum_{n=0}^{\infty} q(n_0 + 2n) \frac{(n_0 + 2n)!}{n_0! n! n!} (\eta^{\text{in}}_{10})^n (\eta^{\text{in}}_{01})^n \eta_{11}^{n_0},$$
$$\eta_{10} = \sum_{n0=0}^{\infty} \sum_{n_1 > n_2}^{\infty} q(n_0 + n_1 + n_2) \frac{(n_0 + n_1 + n_2)!}{n_0! n_1! n_2!} \times (\eta^{\text{in}}_{10})^{n_1} (\eta^{\text{in}}_{01})^{n_2} \eta_{11}^{n_0}. \quad (7)$$

Together with the normalization condition $1 = \eta_{10} + \eta_{01} + \eta_{11}$ this forms a closed set of equations. We have used the abbreviations $\eta^{\text{in}}_{10} = p_{\text{in}}\eta_{10} + (1 - p_{\text{in}})\eta_{01}$ and $\eta^{\text{in}}_{01} = p_{\text{in}}\eta_{01} + (1 - p_{\text{in}})\eta_{10}$. Equations (7) are easily solved for any value of p_{in} and any degree distribution $p(k)$ by iteration. For $p_{\text{in}} = 1/2$, we must have $\eta_{10} = \eta_{01} = \eta_1$ and only one independent order parameter remains.

The ground state energy of the partitioning problem is given by

$$E = -\frac{\langle k \rangle}{2}(1 + 2(X - \eta_{10}\eta_{01}) - (1 - p_{\text{in}})(\eta_{10} - \eta_{01})^2), \quad (8)$$

where we have introduced X as an abbreviation for

$$X = \frac{1}{\langle k \rangle} \sum_{n_0=0}^{\infty} \sum_{n=1}^{\infty} p(n_0 + 2n) \frac{(n_0 + 2n)!}{n_0! n! (n-1)!} \eta_{11}^{n_0} (\eta^{\text{in}}_{10})^n (\eta^{\text{in}}_{01})^n. \quad (9)$$

In case of a Poissonian degree distribution $p(k) = e^{-\lambda}\lambda^k/k!$ with mean λ, we can express this using Modified Bessel Functions of the first kind $I_1(n, x)$:

$$X_\lambda = \sqrt{\eta^{\text{in}}_{10}\eta^{\text{in}}_{01}}\, e^{-\lambda(1-\eta_{11})} I_1(1, 2\lambda\sqrt{\eta^{\text{in}}_{10}\eta^{\text{in}}_{01}}). \quad (10)$$

We denote the ground state energy for $p_{\text{in}} = 1/q$ by E^{Rnd} in which case we have $\eta_{10} = \eta_{01}$.

Once we have the distributions $Q^s(\mathbf{u})$ and hence $P^s(\mathbf{h})$, we can calculate the σ_i conditional on the hidden variables s_i. Node i is assigned state σ_i corresponding to the maxi-

mum component of the effective field $\mathbf{h}$ which is distributed as $P^s(\mathbf{h})$. In case of degeneracy, σ_i is chosen with equal probability among the different maximum components. From this the accuracy follows.

Figure 1 shows order parameters, ground state energy, and achievable accuracy of recovering a planted bisection as a function of p_{in} in a random Bethe lattice with exactly three links per node. The order parameters η_{10} and η_{01}, i.e., the probabilities of sending a message indicating the correct or wrong cluster, respectively, are equal until a critical value of p_{in}^c is reached. This bifurcation of the order parameter pair $\eta_{1,w-1}$, $\eta_{0,w}$ is also observed for more than two clusters. The ground state energy is equal to E^{Rnd} as long as $p_{\text{in}} < p_{\text{in}}^c$. The ground state configuration has only random overlap with the planted partition until $p_{\text{in}} > p_{\text{in}}^c$. As long as $p_{\text{in}} < p_{\text{in}}^c$, the planted partition does not influence the ground state and is thus not detectable. The value of $p_{\text{in}}^c = 7/8$ at which the planted solution starts to influence the ground state is smaller than the naïve guess $p_{\text{in}}^n = -2E^{\text{Rnd}}/\langle k\rangle = 25/27$, the value for which the planted solution starts to have an energy below $E^{\text{Rnd}} = -25/18$. We also see that the accuracy rises quickly as soon as E is lower than E^{Rnd}.

How does the critical value p_{in}^c change with the degree distribution? At the transition point, we can set $\eta_{10} = \eta_{01} + \delta \approx \eta_1$. Then we have $\eta_{10}^{\text{in}} = \eta_{10} - \delta p_{\text{out}}$ and $\eta_{01}^{\text{in}} = \eta_{10} - \delta p_{\text{in}}$. Inserting this into (7) and expanding for small δ we arrive at

$$(p_{\text{in}}^c - p_{\text{out}}^c)^{-1} = \sum_{n_0=0} \sum_{n_1>n_2} q(n_0+n_1+n_2)(n_1-n_2) \times \frac{(n_0+n_1+n_2)!}{n_0!n_1!n_2!} \eta_1^{n_1+n_2-1} \eta_2^{n_0}. \quad (11)$$

Here, η_1 and η_2 are the order parameters that we calculate for $p_{\text{in}} = 1/2$ and that remain valid for all $p_{\text{in}} \le p_{\text{in}}^c$. Expression (11) is easily evaluated for any degree distribution. For a Poissonian degree distribution $p(k) = e^{-\lambda}\lambda^k/k!$ with mean λ, it simplifies to

$$(p_{\text{in}}^c - p_{\text{out}}^c)^{-1} = \lambda(\eta_2 + X_\lambda/\eta_1). \quad (12)$$

Figure 2 shows the dependence of p_{in}^c on the degree distribution. With increasing $\langle k\rangle$ we find decreasing p_{in}^c. However, the critical p_{in} for distributions with fat tails is lower than for networks with a Poissonian degree distribution. Note the correspondence to the results in Ref. [13] on the cut-size of these graphs. The critical value of p_{in} is smaller, i.e., clusters are easier to detect, for networks with degree distributions which are harder to cut. Ref. [13] suggests a universal dependence of E^{Rnd} on $\langle\sqrt{k}\rangle$ based on a replica calculation. Our calculations here support this result. The middle panel of Fig. 2 shows that the naïve estimate $p_{\text{in}}^c \approx p_{\text{in}}^n = -2E^{\text{Rnd}}/\langle k\rangle$ provides a good, but conservative, approximation for large $\langle k\rangle$.

All the results described here analytically for two clusters can be obtained for more than two clusters by an efficient population dynamics algorithm described elsewhere [14]. For example, the right panel of Fig. 2 shows the maximum attainable accuracy in a commonly used benchmark in graph clustering or community detection [15]. Networks consist of 4 groups, the degree distribution is Poissonian with a mean degree of $\lambda = 16$. As expected, our theoretical curve is nicely approached by the data points as the networks grow in size.

In summary, motivated by numerical evidence for a universal limit of cluster detectability across a variety of algorithms [15], we have shown analytically that the sparsity of a network limits the achievable accuracy. Cluster structure may be present, but remains undetectable and hidden behind alternative solutions to the clustering problem that have zero correlation with the correct solution. We have given analytical formulas for the energy of these alternative solutions. To be detectable, any planted configuration must induce lower minima in the energy landscape of the partitioning Hamiltonian. While presented for equal sized clusters, the observed transitions occur in the same qualitative manner for more general cluster structures [14].

This is in strong contrast to a rich literature on clustering multivariate data. The typical behavior of these problems is

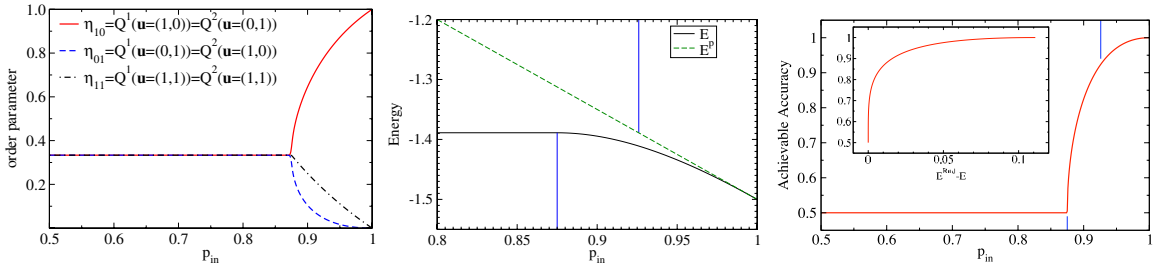

FIG. 1 (color online). Left: Order parameters η_{cw} for the planted bisection on a random Bethe lattice with $k = 3$ links per node as a function of p_{in}. The clusters do not influence the ground state configuration until a critical value of p_{in} is reached. Middle: Ground state energy E of (1) and the energy of the planted cluster structure E^p vs p_{in}. The left vertical line indicates the critical value of p_{in}^c beyond which $\eta_{10} > \eta_{01}$, $E < E^{\text{Rnd}}$, and the clusters do influence the ground state. The right vertical line indicates the naïve guess for $p_{\text{in}}^n = -2E^{\text{Rnd}}/\langle k\rangle$ beyond which $E^p < E^{\text{Rnd}}$. Right: The achievable accuracy for recovering the planted clusters. The two vertical lines indicate p_{in}^c and p_{in}^n. The inset shows how dramatically the accuracy increases for $E < E^{\text{Rnd}}$.

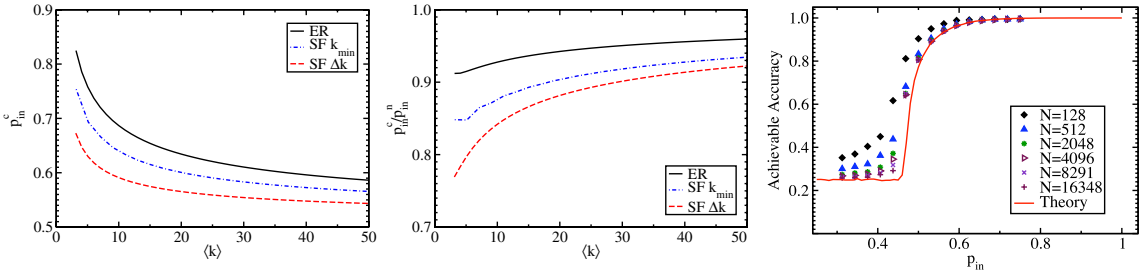

FIG. 2 (color online). Left: The critical value of p_{in} beyond which the cluster structure starts to influence the ground state of the bisection problem, i.e., below which clusters cannot be detected. We compare Erdös Renyi graphs (ER) with a Poissonian degree distribution $p(k) = e^{-\lambda}\lambda^k/k!$ and two types of scale free degree distributions. The first one being a stretched power law (SF Δk) of form $p(k) = (k + \Delta k)^{-\gamma}$ with $\Delta k \in [2, 50]$, and the second (SF k_{min}) being of the form $p(k) = k^{-\gamma}$ with a varying minimum degree k_{min} with $k_{\text{min}} \in [2, 30]$. For both scale free distributions we choose $\gamma = 3$. Since we are interested only in the behavior of the giant connected component, we set $p(k = 0) = 0$ in all cases. Middle: The ratio of p_{in}^c and p_{in}^n. The naïve estimate for the transition point $p_{\text{in}}^n = -2E^{\text{Rnd}}/\langle k\rangle$ always overestimates the true p_{in}^c. Right: Achievable accuracy for the planted partition problem on ER graphs with $N \to \infty$, $\langle k\rangle = 16$ and 4 equal sized clusters and numerical results obtained for corresponding finite size test-networks of varying size. Partitioning was done by simulated annealing.

that given N data points in a space of dimension D, i.e., an $N \times D$ data matrix, there exists a critical value of α_c, such that for $N > \alpha_c D$ one can recover the cluster structure in the data with high accuracy [1,16–19]. Naturally, α_c is a function of the separation of the clusters, but given enough data points, even for smallest separations we are always able to infer the correct cluster structure. A similar result has been derived in the computer science literature by Onsjö and Watanabe for dense networks. They provide an algorithmic proof that a cluster structure can be recovered correctly with probability greater than $1 - \delta$ if $p - r > \Omega(N^{-1/2}\log(N/\delta))$ [20]. In contrast to our treatment, they denote by p and r the probabilities that a link *exists* between nodes in the same, respectively different clusters. Similar bounds are provided by other authors [21,22]. Such bounds are only meaningful if p and r do not scale with the system size. For the sparse networks considered here, however, these bound are meaningless since both p and r scale as $1/N$.

The interesting feature of sparse networks is that size and dimensionality of the data set are not independent. Adding nodes to the network inevitably increases the dimensionality of the data. Thus we are dealing with a qualitatively different phenomenon. Our results may be valuable for the design of network clustering algorithms and their benchmarks as well as for a critical assessment of the amount of information that can be derived from unsupervised learning or data mining on networks.

We thank David Saad, Wolfgang Kinzel, and Georg Reents for stimulating discussions.

[1] A. Engel and C. V. den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, England, 2001).
[2] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[3] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[4] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[5] Y. Fu and P. W. Anderson, J. Phys. A **19**, 1605 (1986).
[6] I. Kanter and H. Sompolinsky, J. Phys. A **20**, L636 (1987).
[7] M. J. de Oliveira, J. Stat. Phys. **54**, 477 (1989).
[8] P.-Y. Lai and Y. Y. Goldschmidt, J. Stat. Phys. **48**, 513 (1987).
[9] W. Liao, Phys. Rev. Lett. **59**, 1625 (1987).
[10] M. Mezard and G. Parisi, J. Stat. Phys. **111**, 1 (2003).
[11] A. Braunstein, R. Mulet, A. Pagnani, M. Weigt, and R. Zecchina, Phys. Rev. E **68**, 036702 (2003).
[12] A. Vázquez and M. Weigt, Phys. Rev. E **67**, 027101 (2003).
[13] J. Reichardt and S. Bornholdt, Phys. Rev. E **76**, 015102(R) (2007).
[14] J. Reichardt and M. Leone (to be published).
[15] L. Danon, J. Dutch, A. Arenas, and A. Diaz-Guilera, J. Stat. Mech. (2005) P09008.
[16] M. Biehl and A. Mietzner, Europhys. Lett. **24**, 421 (1993).
[17] C. V. den Broeck and P. Reimann, Phys. Rev. Lett. **76**, 2188 (1996).
[18] P. Reimann and C. V. den Broeck, Phys. Rev. E **53**, 3989 (1996).
[19] A. Buhot and M. B. Gordon, Phys. Rev. E **57**, 3326 (1998).
[20] M. Onsjö and O. Watanabe, in *ISAAC 2006*, edited by T. Asano (Springer-Verlag, Berlin, Heidelberg, 2006), no. 4288 in LNCS, p. 507.
[21] A. Condon and R. M. Karp, in *Random-Approx'99*, edited by D. Hochbaum (Springer-Verlag, Berlin, Heidelberg, 1999), no. 1671 in LNCS, p. 221.
[22] T. Carson and R. Impagliazzo, in *SODA '01: Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms* (Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001), p. 903.