Community detection and the stochastic block model

Emmanuel Abbe*

February 20, 2016

Abstract

This note surveys some of the recent developments on community detection and the stochastic block model. It describes the fundamental limits of community detection for various recovery requirements, the connections with information theory, and some of the algorithms that emerged in the quest of the thresholds. A few open problems are also discussed. Part of this material was covered in our ISIT 2015 Tutorial with M. Wainwright on Information Theory and Machine Learning.

1 Introduction

The basic task of community detection (or clustering) consists in partitioning the vertices of a graph into clusters that are more densely connected. More generally, community structures may also refer to groups of vertices that connect similarly to the rest of the graphs without having necessarily a higher inner density. In particular, diassortative communities refer to clusters that have higher external connectivity, in contrast to assortative communities. In addition, community detection may be performed on graphs where edges have labels, intensities, or hyper-edges, and communities may not always be well separated, due to overlaps. In the most general context, community detection refers to the problem of inferring similarity relationships among the items of a network by observing their local interactions.

Community detection is one of the central problems in network and data sciences. Virtually any data sets can be represented as a network of interacting items, and one of the first features of interest in such networks is to understand which items are "alike," i.e., communities. Solving this task reliably can provide major insight on understanding sociological behavior [For10, NWS], protein to protein interactions [CY06, MPN+99], gene expressions [CSC+07, JTZ04], recommendation systems [LSY03, SC11], medical prognosis [SPT+01], DNA folding [CAT15], image segmentation [SM97] and the list goes on.

While the field of community detection (CD) has been expanding greatly since the 80's, with impressive developments at the algorithmic and application level, a major part of it has remained for long more an art than a science. In particular, understanding which structures can be extracted, or which are artefacts of algorithms, or how accurate a given clustering may be, are far from being resolved.

*Program in Applied and Computational Mathematics, and Department of Electrical Engineering, Princeton University, Princeton, USA, eabbe@princeton.edu, www.princeton.edu/∼eabbe

The stochastic block model (SBM) has been used widely as a canonical model to study these questions. The SBM is arguably the simplest model of a graph with communities (see definitions in the next section), but like the discrete memoryless channel in coding theory, it provides already strong insights. In addition, the SBM has recently turned into more than a model for community detection. It provides generally a fertile ground for studying various central questions in machine learning, computer science and statistics: It is rich in phase transitions [DKMZ11, Mas14, MNS14b, ABH15, AS15b], allowing to study the interplay between statistical and computational barriers [YC14, AS15c], as well as the discrepancies between probabilstic and adversarial models [MPW15], it serves as an ideal test bed for algorithms, such as SDPs [ABH15, BH14, GV14, AL14, MS15], spectral methods [Vu14, Mas14, BLM15, YP14], belief propagation [KMM$^+$13, AS15c], and it creates new synergies between statistical physics, discrete probability and information theory.

In the next section, we define the SBM and various recovery requirements that are studied for community detection, namely weak, partial and exact recovery. We then provide in Section 3 recent results that have established the fundamental limits for these recovery requirements. We further discuss in Section 4 the connections between information theory and community detection, and give a list of open problems in Section 5.

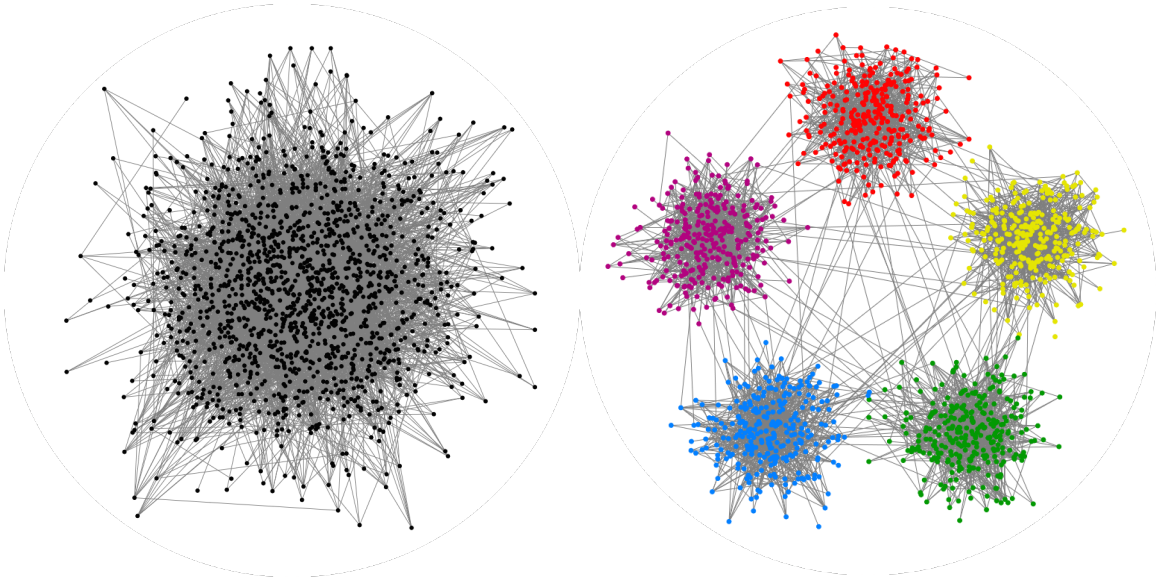


Figure 1: The above two graphs are the same graph re-organized and drawn from the SBM model with 1000 vertices, 5 balanced communities, within-cluster probability of 1/50 and across-cluster probability of 1/1000. The goal of community detection is to obtain the right graph (with the true communities) from the left graph (scrambled) up to some level of accuracy.

2 The stochastic block model

The stochastic block model (SBM) is widely employed as a canonical model for community detection. The history of the SBM is long, and we omit a comprehensive treatment here. Interestingly, the model appeared independently in multiple scientific communities: the terminology SBM, which seems to have dominated in the recent years, comes from the machine learning and statistics literature [HLL83], while the model is typically called the planted partition model in theoretical computer science [BCLS87, DF89, Bop87], and the inhomogeneous random graphs model in the mathematics literature [BJR07].

Definition 1. *Let n be a positive integer (the number of vertices), $p = (p_1, \dots, p_k)$ be a probability vector on $[k] := \{1, \dots, k\}$ (the relative sizes of the communities) and W be a $k \times k$ symmetric matrix with positive entries (the connectivity probabilities). The pair (X, G) is drawn under $\mathrm{SBM}(n, p, W)$, if X is an n-dimensional random vector with components valued in $[k]$ in proportions p (this means that X is either drawn uniformly at random with $\frac{1}{n}|\{v \in [n] : X_v = i\}| = p_i + o(1)$, or with i.i.d. components under p), and G is an n-vertex undirected graph where vertices i and j are connected with probability W_{X_i, X_j}, independently of other pairs of vertices.*

The goal of community detection is to recover the labels X by observing G, up to some level of accuracy.

Definition 2. *(i) The agreement between two community vectors $x, \hat{x} \in [k]^n$ is obtained by minimizing the Hamming distance between x and any relabelling of $\hat{x}$, i.e., any transformation of the components of $\hat{x}$ with a fixed permutation of $[k]$.*
(ii) An algorithm detects communities with accuracy $\alpha \in [0, 1]$, if it takes G drawn from $\mathrm{SBM}(n, p, W)$ and outputs a reconstruction $\hat{X}$ of X that has agreement α with probability $1 - o_n(1)$.

Note that the relabelling in first item above is needed to handle symmetric communities (see below), as it is impossible to recover the actual labels in this case, in contrast to the *partition* which is the object of interest. We now define specific recovery requirements.

Definition 3. *(i) Exact recovery is solvable in $\mathrm{SBM}(n, p, W)$ if there exists an algorithms with accuracy $\alpha = 1$. (ii) Strong recovery is solvable in $\mathrm{SBM}(n, p, W)$ if there exists an algorithms with accuracy $\alpha = 1 - o_n(1)$. (iii) Weak recovery (or detection) is solvable in $\mathrm{SBM}(n, u, V)$, where u is the uniform distribution on $[k]$ and V has constant value in and outside the diagonal, if there exists an algorithms with accuracy $\alpha = 1/k + \varepsilon$ for some $\varepsilon > 0$.*

In other words, exact recovery requires perfect reconstruction of the communities, strong recovery requires almost perfect reconstruction, and weak recovery requires to improve on what a random guess would provide (i.e., $1/k + o(1)$). The most general problem is to understand which accuracy $\alpha \in (0, 1)$ can be achieved in terms of the parameters p and W.

3 Results

3.1 Strong and exact recovery

Exact recovery for linear size communities has long been studied for the SBM [BCLS87, DF89, Bop87, SN97, CK99, McS01, BC09, CWA12, Vu14, YC14], but it is only in the recent years that the fundamental limits were obtained [ABH15, MNS14a, AS15b]. Note that exact recovery requires the node degrees to be at least logarithmic; to see this, note that in the symmetric SBM with disconnected clusters, exact recovery amounts to ask for connectivity in the Erdős-Rényi model, which has a phase transition in the logarithmic degree regime [ER60]. Interestingly, exact recovery has also a phase transition that extends the connectivity one, and is governed by an f-divergence reminiscent of Shannon's coding theorem:

Theorem 1. *[AS15b] Exact recovery is solvable in* $\mathrm{SBM}(n, p, \log(n)Q/n)$ *if and only if*

$$J(p,Q) := \min_{1\le i<j\le k} D_+((\mathrm{diag}(p)Q)_i \| (\mathrm{diag}(p)Q)_j) \ge 1$$

where D_+ *is defined by*

$$D_+(\mu\|\nu) = \max_{t\in[0,1]} \sum_x \nu(x) f_t(\mu(x)/\nu(x)), \quad f_t(y) = 1 - t + ty - y^t, \tag{1}$$

Further, the threshold is efficiently achievable.

Theorem 1 gives an operational meaning to a new f-divergence, D_+, which we call the CH-divergenge in [AS15b] as it generalizes both the Chernoff and Hellinger (or Rényi) divergences. The fundamental limit for data clustering in SBMs is hence governed by the CH-divergence, similarly to the fundamental limit for data transmission in DMCs governed by the KL-divergence. If the columns of $\mathrm{diag}(p)Q$ are "different" enough, where difference is measured in D_+, then one can separate the communities. This is analog to the channel coding theorem, showing that when the output's distributions are different enough in KL-divergence, the codewords can be separated.

To prove the converse, namely, that exact recovery is information-theoretically impossible if $J(p,Q) < 1$, we show that Maximum A-Posteriori (MAP) decoding fails if there exist $i \neq j$ such that $D_+((\mathrm{diag}(p)Q)_i \| (\mathrm{diag}(p)Q)_j) < 1$. This is shown using the following reduction to a *genie-aided community detection problem.* Assume that a genie reveals all the vertices' labels except for a single vertex $v \in [n]$. Then classifying v requires solving an hypothesis test between the k hypotheses corresponding to the k communities, based on the connections that v has with each of the k communities. Here our key result is that, if d_v denotes the vector whose i-th component gives the number of neighbors that v has in community i, d_v is a approximately a multi-variate Poisson random vector with mean $(\mathrm{diag}(p)Q)_{X_v}$ and covariance $\mathrm{diag}((\mathrm{diag}(p)Q)_{X_v})$, and MAP decoding fails with probability roughly given by

$$n^{-\min_{i<j} D_+((\mathrm{diag}(p)Q)_i \| (\mathrm{diag}(p)Q)_j)}. \tag{2}$$

While this is vanishing, it does so too slowly if $D_+((\mathrm{diag}(p)Q)_i \| (\mathrm{diag}(p)Q)_j) < 1$ to prevent that at least one of the $\Theta(n)$ vertices in communities i or j gets misclassified with such a genie-aided test, and thus the non-genie-aided MAP decoder also fails in that case.

To prove that $J(p,Q) > 1$ is an achievable region (and efficiently achievable), we use an efficient algorithm based on a two-round procedure. We start with a "graph-splitting", i.e., we split our original graph into two subgraphs (complement to each other, but essentially independent due to the sparsity of the original graph). On the first graph, whose average degree is taken to be diverging by sub-logarithmic, we run an algorithm that obtains strong recovery. We refer to the next section for the type of algorithms that allows to achieve this. This shows in particular that strong recovery is achievable efficiently as long as the average degrees of the vertices are diverging, i.e., $W = \omega(1)Q/n$. Then we enhance this preliminary clustering by using the left over graph, "cleaning up" the strong clustering into an exact clustering with local improvements based on the hypothesis test described above. If $D_+((\mathrm{diag}(p)Q)_i \| (\mathrm{diag}(p)Q)_j) > 1$ for all $i < j$, the vertices that were misclassified in part 1 can be re-classified correctly with high probability, even though our genie gives now only an approximate clustering. Further, if the algorithm for part 1 is efficient, the whole algorithm is efficient since the local improvement part is only linear in n. Our algorithm 'degree-profiling' has in fact an overall complexity of $O(n^{1+\varepsilon})$, for any $\varepsilon > 0$.

3.2 Weak recovery

Weak recovery was introduced in [Co10, DKMZ11]. Note that weak recovery is investigated in SBMs where vertices have constant expected degree, as otherwise the problem can easily be solved by exploiting the degree variations. The following conjecture was established first in [DKMZ11] from deep but non-rigorous statistical physics arguments, and is responsible in part for the resurged interest in the fundamental study of the SBM:

Conjecture 1. *[DKMZ11, MNS12] Denote by* $\mathrm{SBM}(n,k,a,b)$ *the symmetric sparse SBM, i.e., the model* $\mathrm{SBM}(n,p,W)$ *where* p *is uniform on* $[k]$ *and* $W_{i,j}$ *is* a/n *if* $i = j$ *and* b/n *otherwise. Define* $\mathrm{SNR} = \frac{(a-b)^2}{k(a+(k-1)b)}$*, then*

(i) irrespective of k*, if* $\mathrm{SNR} > 1$ *(the Kesten-Stigum threshold), it is possible to detect communities in polynomial time;*

(ii) if $k \geq 5$*, it is possible to detect communities information-theoretically for some* SNR *strictly below 1.*

We have recently proved this conjecture in [AS15a]. For the case of $k = 2$, it was already proved in [Mas14, MNS14b] that the KS threshold can be achieved efficiently. However, for $k = 2$, no information-computation gap takes place as shown with a tight converse in [MNS12].

The terminology 'KS threshold' comes from the reconstruction problem on trees. A transmitter broadcasts a uniform bit to some relays, which themselves forward the received bits to other relays, etc. The number of relays (or offspring) at each generation may be a constant c, or Poisson distributed of mean c. Each relay is assumed to transmit with an independent BSC of parameter ε. The receiver gets to see all the bits at the leaves. For what values of c and ε could the receiver reconstruct the original bit when the tree depth diverges? The unorthodox part is that we are interested in recovering the bit weakly, i.e., with probability away from $1/2$, and not tending to 1 as usual in information theory. This problem was first solved in [KS66] for binary symmetric channels and constant offspring,

showing that weak recovery is possible if and only if $c > 1/(1-2\varepsilon)^2$, i.e., the KS threshold. It was later solved for the Poisson case in [EKPS00]. This implies a converse for weak recovery in the 2-community SBM [MNS12], using a genie-aided argument and the fact that a node's neighborhood in the sparse SBM is tree-like (in particular $c = (a+b)/2$, $\varepsilon = b/(a+b)$, and the KS threshold reads $(a-b)^2 > 2(a+b)$).

Note that the KS threshold raises an interesting challenge for community detection algorithms, as standard clustering methods fail to detect communities down the KS threshold. This includes spectral methods based on the adjacency matrix or Laplacians [Co10, KMM+13] or SDPs [MS15]. For standard spectral methods, a first issue is that the fluctuations in the node degrees produce high-degree nodes that disrupt the eigenvectors from concentrating on the clusters. One possibility is to trim such high-degree nodes, throwing away some information, but this does not suffice to get the KS threshold.

The first efficient algorithms that managed to achieve the KS threshold for $k = 2$ were based on counting self-avoiding walks (entry (i, j) counts the number of self-avoiding walks of moderate size between vertices i and j) [Mas14], and weighted non-backtracking walks between vertices [MNS14b]:

Theorem 2. *For $k = 2$,*

1. *[Mas14, MNS14b] Weak recovery is solvable efficiently if* SNR > 1 *(i.e., KS threshold is efficiently achievable for $k = 2$);*

2. *[MNS12] Weak recovery is information-theoretically not solvable if* SNR ≤ 1.

It was also shown in [BLM15] that for SBMs with multiple but slightly asymmetrical communities, the KS threshold can be achieved using a spectral method with the matrix of non-backtracking walks between directed edges (each edge is replaced with two directed edges and entry (e, f) is one if and only if edge e follows edge f) [BLM15]. However, [BLM15] does not resolve Conjecture 1 for $k \geq 3$.

We proved Conjecture 1 for arbitrary k using a message passing algorithm:

Theorem 3. *[AS15a] Conjecture 1 holds for all $k \geq 2$. In particular*

1. *Weak recovery is solvable in $O(n \log n)$ if* SNR > 1 *with Acyclic Belief Propagation (ABP), a belief propagation algorithm that is linearized and exploits cycles;*

2. *Weak recovery is information-theoretically solvable for some* SNR *strictly below* 1 *if $k \geq 5$ with Typicality Sampling, a non-efficient algorithm that samples uniformly at random a clustering having the typical proportion of edges inside and across clusters.*

The fact that BP with a random initialization could achieve the KS threshold for arbitrary k was believed to take place [DKMZ11], but handling random initialization and cycles stood has a challenge. Interestingly, our ABP algorithm is also closely related to the non-backtracking operator from [KMM+13], but it improves on the complexity of spectral methods due to the message-passing implementation.

The information-theoretic (IT) bound is characterized in [AS15a] at the extremal regimes of a and b. For $a = 0$, it is shown that weak recovery is information-theoretically solvable if $b > ck \ln k + o_k(1)$, $c \in [1, 2]$. Thus the information-computation gap — defined as

the gap between the KS threshold and the IT bound — is large since the KS threshold reads $b > k(k-1)$. The behaviour of the IT bound is also characterized for b close to 0. Similar though weaker results were also recently posted in [BM16]. Note also that the information-computation gap concerns the gap between the KS threshold and what is achieved information-theoretically, which is the gap between the information-theoretic and computational thresholds only under non-formal evidences [DKMZ11]. Showing formally that no algorithms can succeed below the KS threshold would naturally require novel techniques and major progress on deep complexity theory questions.

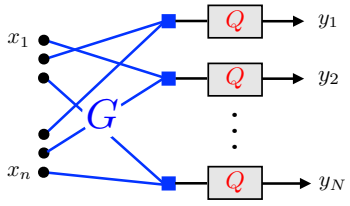
4 Information theory and community detection

Community detection has natural connections with information theory at various levels. Exact recovery is closely related to the decoding of graph-based codes on memoryless channels, and to f-divergences. Weak recovery relates naturally to the broadcasting problem on trees. In the next section, we also mention how partial recovery is connected to information-estimation measures. More generally, community detection pairs well with information theory as it can be viewed as a decoding problem on a noisy channel: the community labels are the input to a black-box channel that provides local and noisy interactions of the inputs. This view point was further developed in [AM15], with the notion of graphical channels:

Definition 4. *[AM15] Let $V = [n]$ and $G = (V, E(G))$ be a hypergraph with $N = |E(G)|$. Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite sets called respectively the input and output alphabets, and $Q(\cdot|\cdot)$ be a channel from $\mathcal{X}^k$ to $\mathcal{Y}$ called the kernel. To each vertex in V, assign a vertex-variable in $\mathcal{X}$, and to each edge in $E(G)$, assign an edge-variable in $\mathcal{Y}$. Let y_I denote the edge-variable attached to edge I, and $x[I]$ denote the k node-variables adjacent to I. We define a graphical channel with graph G and kernel Q as the channel $P(\cdot|\cdot)$ given by*

$$P(y|x) \equiv \prod_{I \in E(G)} Q(y_I|x[I])$$

$$x \in \mathcal{X}^V, y \in \mathcal{Y}^{E(G)}$$



The above departs significantly from a traditionally encoded channel when considering low order edges (e.g., $k = 2, 3$) and G uniform or complete (the closest would be a special LDGM code [KPSS10]). As discussed in [AS15b], exact recovery in the SBM is verbatim a decoding problem on such a channel with an LDGM code of right-degree 2.

Community detection has a strong connection with information theory since $\mathcal{X}$ is typically discrete (as the goal is to obtain 'clusters' on the data), which is not common for other applications in machine learning where the real-valued nature of the channel is important.[1]

[1] Compressed sensing or topic modelling rely instead heavily on real-valued channels.

Graphical channels allow also to capture many extensions of the SBM, such as non-overlapping communities, edge-labeled or non-pairwise interactions. This can be further extended to problems such as topic modelling or ranking, with new notions of recovery. Ubiquitous to all these models are two quantities: a measure on how "rich" the observation graph G is (e.g., the node degrees in the SBM), and a measure on how "noisy" the connectivity kernel Q is (e.g., the CH-divergence for exact recovery). These are not the usual notions of rates and capacity in information theory, but they are the relevant ones here. These also make the problems novel and interesting. By understanding various instances of such models, the hope is to build a general theory for the fundamental limits in machine learning and data science problems, inspired by information theory.

5 Open problems

The establishment of fundamental limits for community detection in the SBM have appeared in the recent years. There is therefore a long list of open problems and directions to pursue, both related to the SBM and to similar models in machine learning. We provide here a partial list:

- *Exact recovery for sub-linear communities.* Theorem 1 gives a comprehensive result for exact recovery in the case of linear-size communities, i.e., when the entries of p and its dimension k do not scale with n. If $k = o(\log(n))$ and the communities are balanced, most of the current techniques apply. However new phenomena seem to take place beyond that, with again gaps between information and computational thresholds. In [YC14], some of this is captured by looking at coarse regimes of the parameters. Finer scale regimes may reveal further interesting directions to explore concerning the discrepancies of information and computation barriers.

- *Partial recovery.* Between weak and exact recovery, *how much* can we hope to recover about the communities? What are the fundamental tradeoffs between the SNR and the distortion/accuracy of detection algorithms? Is there a *rate-distortion theory* of CD? Recently, we were able to answer this question in [DAM15] for the special case where the SNR is constant while the average degrees diverge at the same rate. In particular the mutual information and I-MMSE formula [GSV05] allow to estimate the SNR-distortion curve sharply. The finite SNR regime with constant node degrees, or the case with multiple (asymmetric) communities remain open.

- *The information-computation gap.* Can we locate the exact information-theoretic threshold for weak recovery when $k \geq 3$? Can we strengthen the evidences that the KS threshold is the computational threshold?

- *Beyond the SBM.* How do previous results generalize to other graphical channels [AM15, ABBS14]?

References

[ABBS14] E. Abbe, A.S. Bandeira, A. Bracher, and A. Singer, *Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery*, Network Science and Engineering, IEEE Transactions on **1** (2014), no. 1, 10–22. 8

[ABH15] E. Abbe, A.S. Bandeira, and G. Hall, *Exact recovery in the stochastic block model*, Information Theory, IEEE Transactions on **62** (2015), no. 1, 471–487. 2, 4

[AL14] A. Amini and E. Levina, *On semidefinite relaxations for the block model*, arXiv:1406.5647 (2014). 2

[AM15] Emmanuel Abbe and Andrea Montanari, *Conditional random fields, planted constraint satisfaction, and entropy concentration*, Theory of Computing **11** (2015), no. 17, 413–443. 7, 8

[AS15a] E. Abbe and C. Sandon, *Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic BP, and the information-computation gap*, ArXiv e-prints 1512.09080 (2015). 5, 6

[AS15b] Emmanuel Abbe and Colin Sandon, *Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery*, IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015, 2015, pp. 670–688. 2, 4, 7

[AS15c] Emmanuel Abbe and Colin Sandon, *Recovering communities in the general stochastic block model without knowing the parameters*, Advances in Neural Information Processing Systems (NIPS) 28 (C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett, and R. Garnett, eds.), Curran Associates, Inc., 2015, pp. 676–684. 2

[BC09] P. J. Bickel and A. Chen, *A nonparametric view of network models and newmangirvan and other modularities*, Proceedings of the National Academy of Sciences (2009). 4

[BCLS87] T.N. Bui, S. Chaudhuri, F.T. Leighton, and M. Sipser, *Graph bisection algorithms with good average case behavior*, Combinatorica **7** (1987), no. 2, 171–191. 3, 4

[BH14] J. Xu B. Hajek, Y. Wu, *Achieving exact cluster recovery threshold via semidefinite programming*, arXiv:1412.6156 (2014). 2

[BJR07] Béla Bollobás, Svante Janson, and Oliver Riordan, *The phase transition in inhomogeneous random graphs*, Random Struct. Algorithms **31** (2007), no. 1, 3–122. 3

[BLM15] C. Bordenave, M. Lelarge, and L. Massoulié, *Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs*, Available at arXiv:1501.06087 (2015). 2, 6

[BM16] J. Banks and C. Moore, *Information-theoretic thresholds for community detection in sparse networks*, ArXiv e-prints (2016). 7

[Bop87] R.B. Boppana, *Eigenvalues and graph bisection: An average-case analysis*, In 28th Annual Symposium on Foundations of Computer Science (1987), 280–285. 3, 4

[CAT15] I. Cabreros, E. Abbe, and A. Tsirigos, *Detecting Community Structures in Hi-C Genomic Data*, ArXiv e-prints 1509.05121 (2015). 1

[CK99] A. Condon and R. M. Karp, *Algorithms for graph partitioning on the planted partition model*, Lecture Notes in Computer Science **1671** (1999), 221–232. 4

[Co10] A. Coja-oghlan, *Graph partitioning via adaptive spectral techniques*, Comb. Probab. Comput. **19** (2010), no. 2, 227–284. 5, 6

[CSC+07] M.S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A.R. Pico, A. Vailaya, P. Wang, A. Adler, B.R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G.D. Bader, *Integration of biological networks and gene expression data using cytoscape*, Nature Protocols **2** (2007), no. 10, 2366–2382. 1

[CWA12] D. S. Choi, P. J. Wolfe, and E. M. Airoldi, *Stochastic blockmodels with a growing number of classes*, Biometrika (2012), 1–12. 4

[CY06] J. Chen and B. Yuan, *Detecting functional modules in the yeast proteinprotein interaction network*, Bioinformatics **22** (2006), no. 18, 2283–2290. 1

[DAM15] Y. Deshpande, E. Abbe, and A. Montanari, *Asymptotic mutual information for the two-groups stochastic block model*, arXiv:1507.08685 (2015). 8

[DF89] M.E. Dyer and A.M. Frieze, *The solution of some random NP-hard problems in polynomial expected time*, Journal of Algorithms **10** (1989), no. 4, 451 – 489. 3, 4

[DKMZ11] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications*, Phys. Rev. E **84** (2011), 066106. 2, 5, 6, 7

[EKPS00] W. Evans, C. Kenyon, Y. Peres, and L. J. Schulman, *Broadcasting on trees and the Ising model*, Ann. Appl. Probab. **10** (2000), 410–433. 6

[ER60] P. Erdös and A Rényi, *On the evolution of random graphs*, Publication of the Mathematical Institute of the Hungarian Academy of Sciences, 1960, pp. 17–61. 4

[For10] S. Fortunato, *Community detection in graphs*, Physics Reports **486 (3-5)** (2010), 75–174. 1

[GSV05] Dongning Guo, Shlomo Shamai, and Sergio Verdú, *Mutual information and minimum mean-square error in gaussian channels*, Information Theory, IEEE Transactions on **51** (2005), no. 4, 1261–1282. 8

[GV14] O. Guédon and R. Vershynin, *Community detection in sparse networks via Grothendieck's inequality*, ArXiv:1411.4686 (2014). 2

[HLL83] P. W. Holland, K. Laskey, and S. Leinhardt, *Stochastic blockmodels: First steps*, Social Networks **5** (1983), no. 2, 109–137. 3

[JTZ04] D. Jiang, C. Tang, and A. Zhang, *Cluster analysis for gene expression data: a survey*, Knowledge and Data Engineering, IEEE Transactions on **16** (2004), no. 11, 1370–1386. 1

[KMM+13] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, *Spectral redemption: clustering sparse networks*, CoRR **abs/1306.5550** (2013). 2, 6

[KPSS10] K. R. Kumar, P. Pakzad, A.H. Salavati, and A. Shokrollahi, *Phase transitions for mutual information*, Turbo Codes and Iterative Information Processing (ISTC), 2010 6th International Symposium on, 2010, pp. 137–141. 7

[KS66] H. Kesten and B. P. Stigum, *A limit theorem for multidimensional galton-watson processes*, Ann. Math. Statist. **37** (1966), no. 5, 1211–1223. 5

[LSY03] G. Linden, B. Smith, and J. York, *Amazon.com recommendations: Item-to-item collaborative filtering*, IEEE Internet Computing **7** (2003), no. 1, 76–80. 1

[Mas14] L. Massoulié, *Community detection thresholds and the weak Ramanujan property*, STOC 2014: 46th Annual Symposium on the Theory of Computing (New York, United States), June 2014, pp. 1–10. 2, 5, 6

[McS01] F. McSherry, *Spectral partitioning of random graphs*, Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on, 2001, pp. 529–537. 4

[MNS12] E. Mossel, J. Neeman, and A. Sly, *Stochastic block models and reconstruction*, Available online at arXiv:1202.1499 [math.PR] (2012). 5, 6

[MNS14a] E. Mossel, J. Neeman, and A. Sly, *Consistency thresholds for binary symmetric block models*, Arxiv:arXiv:1407.1591. In proc. of STOC15. (2014). 4

[MNS14b] E. Mossel, J. Neeman, and A. Sly, *A proof of the block model threshold conjecture*, Available online at arXiv:1311.4115 [math.PR] (2014). 2, 5, 6

[MPN+99] E.M. Marcotte, M. Pellegrini, H.-L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg, *Detecting protein function and protein-protein interactions from genome sequences*, Science **285** (1999), no. 5428, 751–753. 1

[MPW15] A. Moitra, W. Perry, and A. S. Wein, *How Robust are Reconstruction Thresholds for Community Detection?*, ArXiv e-prints (2015). 2

[MS15] A. Montanari and S. Sen, *Semidefinite Programs on Sparse Random Graphs and their Application to Community Detection*, ArXiv e-prints (2015). 2, 6

[NWS] M. E. J. Newman, D. J. Watts, and S. H. Strogatz, *Random graph models of social networks*, Proc. Natl. Acad. Sci. USA **99**, 2566–2572. 1

[SC11] S. Sahebi and W. Cohen, *Community-based recommendations: a solution to the cold start problem*, Workshop on Recommender Systems and the Social Web (RSWEB), held in conjunction with ACM RecSys11, October 2011. 1

[SM97] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (1997), 888–905. 1

[SN97] T. A. B. Snijders and K. Nowicki, *Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure*, Journal of Classification **14** (1997), no. 1, 75–100. 4

[SPT+01] T. Sorlie, C.M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, Mi.B. Eisen, M. van de Rijn, S.S. Jeffrey, T. Thorsen, H. Quist, J.C. Matese, P.O. Brown, D. Botstein, P.E. Lonning, and A. Borresen-Dale, *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications*, no. 19, 10869–10874. 1

[Vu14] V. Vu, *A simple svd algorithm for finding hidden partitions*, Available online at arXiv:1404.3918 (2014). 2, 4

[YC14] J. Xu Y. Chen, *Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices*, arXiv:1402.1267 (2014). 2, 4, 8

[YP14] S. Yun and A. Proutiere, *Accurate community detection in the stochastic block model via spectral algorithms*, arXiv:1412.7335 (2014). 2