

Crossing the KS threshold in the stochastic block model with information theory

Emmanuel Abbe
Princeton University
eabbe@princeton.edu

Colin Sandon
Princeton University
sandon@princeton.edu

Abstract—Decelle et al. conjectured that community detection in the symmetric stochastic block model has a computational threshold given by the so-called Kesten-Stigum (KS) threshold, and that information-theoretic methods can cross this threshold for a large enough number of communities (4 or 5 depending on the regime of the parameters). This paper shows that at $k = 5$, it is possible to cross the KS threshold in the disassortative regime with a non-efficient algorithm that samples a clustering having typical cluster volumes. Further, the gap between the KS and information-theoretic threshold is shown to be large in some cases. In the case where edges are drawn only across clusters with an average degree of b , and denoting by k the number of communities, the KS threshold reads $b \gtrsim k^2$ whereas our information-theoretic bound reads $b \gtrsim k \ln(k)$.

I. INTRODUCTION

The stochastic block model (SBM) is a canonical model of network with communities. The terminology SBM comes from the machine learning and statistics literature [1], while the model is typically called the planted partition model in theoretical computer science [2], [3], and the inhomogeneous random graphs model in the mathematical literature [4]. Although the model was defined as far back as the 80s, it resurged in the recent years due in part to the following fascinating conjecture established in [5] (and backed in [10]) from deep but non-rigorous statistical physics arguments:

Conjecture 1. *Let (X, G) be drawn from $SBM(n, k, a, b)$, i.e., X is uniformly drawn among partitions of $[n]$ into k balanced clusters, and G is a random graph on the vertex set $[n]$ where edges are placed independently with probability a/n inside the clusters and b/n across. Define $SNR = \frac{(a-b)^2}{k(a+(k-1)b)}$ and say that an algorithm detects communities if it takes G as input and outputs \hat{X} that is positively correlated with X with high probability. Then,*

- (i) *Irrespective of k , if $SNR > 1$, it is possible to detect communities in polynomial time, i.e., the KS threshold can be achieved efficiently;*

- (ii) *If $k \geq 4$ ($k \geq 5$ in the assortative case), it is possible to detect communities information-theoretically for some SNR strictly below 1.*

We have recently prove part (i) of this conjecture in [6], and this paper shows that for $k = 5$, it is indeed possible to cross the KS threshold using information theory in the disassortative case. For part (i), i.e., achieving the KS threshold efficiently, [6] relies on a linearized version of BP that can handle cycles and runs in $O(n \log n)$ time. This approach is related to spectral methods based on non-backtracking operators [7].

To cross the KS threshold information-theoretically, we rely on a non-efficient algorithm that samples a typical clustering. Upon observing a graph drawn from the SBM, the algorithm builds the set of all partitions of the n nodes that have a typical fraction of edges inside and across clusters, and then samples a partition uniformly at random in that set. Our analysis of this algorithm reveals two different regimes, that reflect two layers of refinement in the bounds on the typical set's size. In a first regime, bad clusterings (i.e., partitions of the nodes that agree in no more than close to $1/k$ vertices) are with high probability not typical using a union-bound, and the algorithm samples only good clusterings with high probability. This allows to cross the KS threshold at $a = 0$, and shows in this case that detection is information-theoretically solvable if $b > ck \ln k + o_k(1)$, $c \in [1, 2]$. Thus the gap between the information-theoretic and KS threshold can be large, since the KS threshold reads $b > k(k-1)$. However, the union bound does not allow to recover the correct bound at $b = 0$. For $b = 0$, previous analysis leads to a bound given by $a > 2k$, which is suboptimal. In fact, as soon as $a > k$, each cluster in the SBM graph has a giant component of linear size, and thus an algorithm that simply separates these components and randomly assigned the remaining vertices will detect the communities. Of course, such an algorithm only applies

to the strict case of $b = 0$. To obtain a tighter bound in the general case, we next exploit the large number of tree-like components that in the SBM graph, reaching a regime where some bad clusterings are typical but unlikely to be sampled. This shows that the algorithm succeed with the right bound¹ at $b = 0$, i.e., $a > k$.

A. Related literature

For the case of $k = 2$, it was proved in [8], [9] that the KS threshold can be achieved efficiently. An alternative proof was later given in [7]. For $k = 2$, no information-computation gap takes places as shown with a tight converse in [10]. It was also shown in [7] that for SBMs with multiple slightly asymmetric communities, the KS threshold can be achieved, but [7] does not resolve Conjecture 1 for $k \geq 3$. Note that the crossing the KS threshold with information theory shows a gap between the information-theoretic and computational thresholds only under non-formal evidences [5]. Note also that standard clustering methods are not believed/known to detect clusters down the KS threshold. This includes spectral methods based on the adjacency matrix or Laplacians or SDPs. 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 achieve the KS threshold [11].

A few papers have studied information-theoretic bounds in SBMs with a growing number of communities [12], two unbalanced communities [13], and a single community [14]. No results seemed so far known for the symmetric SBM and part (ii) of Conjecture 1. Shortly after this paper posting, [15] obtained in an independent effort bounds on the information theoretic threshold that cross the KS threshold at $k = 5$ (in the disassortative case), using moment methods. The bound in [15] does however not approach to the correct threshold at $b = 0$.

II. RESULTS

The SBM can be defined with a uniform or Binomial model for the communities. This means that for a probability vector $p = (p_1, \dots, p_k)$, the communities may be drawn uniformly at random among all partitions of n having np_i vertices in community i (with an arbitrary rounding rule on np_i to obtain integers adding up to n), or each vertex may be assigned a label in $[k]$ independently with probability p . These are equivalent for

¹Further improvements can be obtained from the second regime, with a finer estimate on the typical set's size that exploits also the parts of the giant that are tree-like; see [6]

the purpose of this paper, due to standard concentration argument. In the case where $p = (1/k, \dots, 1/k)$, we say that the communities are balanced.

Definition 1. (X, G) is drawn under $\text{SBM}(n, k, a, b)$, if X is a balanced n -dimensional vector with components valued in $[k]$ and G is a random graph on the vertex set $[n]$ where edge $(i, j) \in \binom{[n]}{2}$ is drawn with probability $1(X_i = X_j)a/n + 1(X_i \neq X_j)b/n$, independently of the other edges.

Note that we often talk about G being drawn under the SBM without specifying the community variables X .

Definition 2. Let $x \in [k]^n$ and $\varepsilon > 0$. We define the set of bad clusterings with respect to x by $B_\varepsilon(x) = \{y \in [n]^k : \frac{1}{n}d_*(x, y) > 1 - \frac{1}{k} - \varepsilon\}$, where $d_*(x, y)$ is the minimum Hamming distance between x and any relabelling of y (i.e., any mapping of the components of y with a fixed permutation of $[k]$).

Relabellings need to be considered since only the partition needs to be detected and not the actual labels. It is simply convenient to work with labels.

Definition 3. An algorithm $\hat{x} : 2^{\binom{[n]}{2}} \rightarrow [k]^n$ solves detection (or weak recovery) in $\text{SBM}(n, k, a, b)$ if for some $\varepsilon > 0$, $\mathbb{P}_{X, G} \{\hat{x}(G) \in B_\varepsilon(X)\} = o_n(1)$, where $(X, G) \sim \text{SBM}(n, k, a, b)$. Detection is solvable efficiently if the algorithm runs in polynomial time in n , and information-theoretically otherwise.

Note that if \hat{X} is a randomized algorithm (i.e., it takes the graph as an input and outputs various clusterings with different probabilities), and if for some $\varepsilon > 0$, $\mathbb{P}_{X, G, \hat{X}} \{\hat{X}(G) \in B_\varepsilon(X)\} = o_n(1)$, then detection is solvable (information-theoretically).

We next present the algorithm used to detect below the KS threshold.

Typicality Sampling Algorithm. Given an n -vertex graph g and $\delta > 0$, the algorithm draws $\hat{x}_{\text{typ}}(g)$ uniformly at random in

$$\begin{aligned} T_\delta(g) &= \{x \in \text{Balanced}(n, k) : \\ &\sum_{i=1}^k |\{g_{u,v} : (u, v) \in \binom{[n]}{2} \text{ s.t. } x_u = i, x_v = i\}| \\ &\geq \frac{an}{2k}(1 - \delta), \\ &\sum_{i, j \in [k], i < j} |\{g_{u,v} : (u, v) \in \binom{[n]}{2} \text{ s.t. } x_u = i, x_v = j\}| \\ &\leq \frac{bn(k-1)}{2k}(1 + \delta)\} \end{aligned}$$

if the SBM is assortative (i.e., $a \geq b$), otherwise flip the direction of the above two inequalities.

Theorem 1. Let $d := \frac{a+(k-1)b}{k}$, assume $d > 1$, and let $\tau = \tau_d$ be the unique solution in $(0, 1)$ of $\tau e^{-\tau} = d e^{-d}$, i.e., $\tau := \sum_{j=1}^{+\infty} \frac{j^{j-1}}{j!} (d e^{-d})^j$. The Typicality Sampling Algorithm detects² communities in $\text{SBM}(n, k, a, b)$ if

$$\frac{1}{2 \ln k} \left(\frac{a \ln a + (k-1)b \ln b}{k} - d \ln d \right) > 1 - \frac{\tau}{d} \left(1 - \frac{\tau}{2} \right). \quad (1)$$

Remark 1. Define $d(\tau, d) = 1 - \frac{\tau}{d} \left(1 - \frac{\tau}{2} \right)$. Note that since $d(\tau, d) < 1$ when $d > 1$ (which is needed for the presence of the giant), detection is already solvable in $\text{SBM}(n, k, a, b)$ if

$$\frac{1}{2 \ln k} \left(\frac{a \ln a + (k-1)b \ln b}{k} - d \ln d \right) > 1. \quad (2)$$

The above corresponds to the regime where there is not bad clustering that is typical with high probability. However, the above bound is not tight in the extreme regime of $b = 0$, since it reads $a > 2k$ as opposed to $a > k$ (the presence of a giant).

Defining $a_k(b)$ as the solution in a of $\frac{1}{2 \ln k} \left(\frac{a \ln a + (k-1)b \ln b}{k} - d \ln d \right) = d(\tau, d)$ and expanding the bound in Theorem 1 gives the following.

Corollary 1. Detection is solvable

$$\text{in } \text{SBM}(n, k, 0, b) \text{ if } b > \frac{2k \ln k}{(k-1) \ln \frac{k}{k-1}} g\left(\tau, \frac{b(k-1)}{k}\right) \quad (3)$$

$$\text{in } \text{SBM}(n, k, a, b) \text{ if } a > k + \Delta_k(b), \quad (4)$$

where (3) is strictly stronger than the KS threshold at $k = 5$, and where $\Delta_k(b) := a_k(b) - k$ is such that $\Delta_k(0) = 0$.

Remark 2. Note that (4) approaches the optimal bound given by the presence of the giant at $b = 0$. Note also that (3) improves significantly on the KS threshold given by $b > k(k-1)$ at $a = 0$. By continuity arguments, we can also cross the KS threshold for some positive values of a at $k = 5$.

Remark 3. We further claim that the scaling in k of our IT threshold is correct $a = 0$. To see this, consider $v \in G$, $b = (1 - \epsilon)k \ln(k)$, and assume that we know the communities of all vertices more than $r = \ln(\ln(n))$

²Setting $\delta > 0$ small enough gives the existence of $\epsilon > 0$ for detection.

edges away from v . For each vertex r edges away from v , there will be approximately k^ϵ communities that it has no neighbors in. Then vertices $r-1$ edges away from v have approximately $k^\epsilon \ln(k)$ neighbors that are potentially in each community, with approximately $\ln(k)$ fewer neighbors suspected of being in its community than in the average other community. At that point, the noise has mostly drowned out the signal and our confidence that we know anything about the vertices' communities continues to degrade with each successive step towards v . A different approach is developed in [15].

III. PROOF TECHNIQUE

A first question is to estimate the likelihood that a bad clustering, i.e., one that has an overlap that is close to $1/k$, belongs to the typical set. This means the probability that a clustering which splits each of the true cluster into k groups belonging to each community still manages to keep the right proportions of edges inside and across the clusters. This is unlikely to take place, but we care about the exponent of this rare event probability.

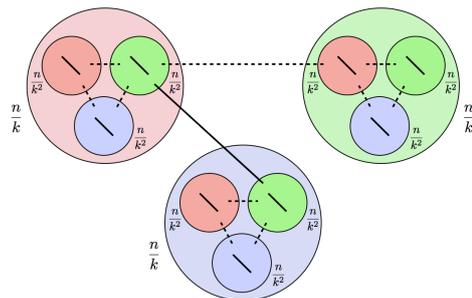


Fig. 1: A bad clustering roughly splits each community equally among the k communities. Each pair of nodes connects with probability a/n among vertices of same communities (i.e., same color groups, plain line connections), and b/n across communities (i.e., different color groups, dashed line connections). Only some connections are displayed in the Figure to ease the visualization.

As illustrated in Figure 1, the number of edges that are contained in the clusters of a bad clustering is roughly distributed as the sum of two Binomial random variables,

$$E_{\text{in}} \dot{\sim} \text{Bin} \left(\frac{n^2}{2k^2}, \frac{a}{n} \right) + \text{Bin} \left(\frac{(k-1)n^2}{2k^2}, \frac{b}{n} \right),$$

where we use $\dot{\sim}$ to emphasize that this is an approximation. Note that the expectation of the above distribution is

$\frac{n}{2k} \frac{a+(k-1)b}{k}$. In contrast, the true clustering would have a distribution given by $\text{Bin}(\frac{n^2}{2k}, \frac{a}{n})$, which would give an expectation of $\frac{an}{2k}$. In turn, the number of edges that are crossing the clusters of a bad clustering is roughly distributed as

$$E_{\text{out}} \sim \text{Bin}\left(\frac{n^2(k-1)}{2k^2}, \frac{a}{n}\right) + \text{Bin}\left(\frac{n^2(k-1)^2}{2k^2}, \frac{b}{n}\right),$$

which has an expectation of $\frac{n(k-1)}{2k} \frac{a+(k-1)b}{k}$. In contrast, the true clustering would have the above replaced by $\text{Bin}(\frac{n^2(k-1)}{2k}, \frac{b}{n})$, and an expectation of $\frac{bn(k-1)}{2k}$.

Thus, we need to estimate the rare event that the Binomial sum deviates from its expectations. While there is a large list of bounds on Binomial tail events, the number of trials here is quadratic in n and the success bias decays linearly in n , which require particular care to ensure tight bounds. We derive these in [6], obtaining that for a bad clustering x ,

$$\mathbb{P}\{x \text{ is typical}\} \approx \exp\left(-\frac{n}{k}A\right)$$

where

$$A := \frac{a+b(k-1)}{2} \ln \frac{k}{a+(k-1)b} + \frac{a}{2} \ln a + \frac{b(k-1)}{2} \ln b.$$

One can then use a union bound, since there are at most k^n bad clusterings, to obtain a first regime where no clustering is typical with high probability. This already allows to cross the KS threshold in some regime of the parameters when $k \geq 5$. However, this does not interpolate the correct behavior of the information-theoretic bound in the extreme regime of $b = 0$. In fact, for $b = 0$, the union bound requires $a > 2k$ to imply no bad typical clustering with high probability, whereas as soon as $a > k$, an algorithm that simply separates the two giants in $\text{SBM}(n, k, a, 0)$ and assigns communities uniformly at random for the other vertices solves detection. Thus when $a \in (k, 2k]$, the union bound is loose. To remediate to this, we next take into account the topology of the SBM graph.

Since the algorithm samples a typical clustering, we only need the number of bad and typical clusterings to be small compared to the total number of typical clusterings, in expectation. Thus, we seek to better estimate the total number of typical clusterings. The main topological property of the SBM graph that we exploit is the large fraction of nodes that are in tree-like components outside of the giant. Conditioned on being on a tree, the SBM labels are distributed as in a broadcasting problem on a (Galton-Watson) tree. Specifically, for a uniformly drawn

root node X , each edge in the tree acts as a k -ary symmetric channel. Thus, labelling the nodes in the trees according to the above distribution and freezing the giant to the correct labels leads to a typical clustering with high probability.

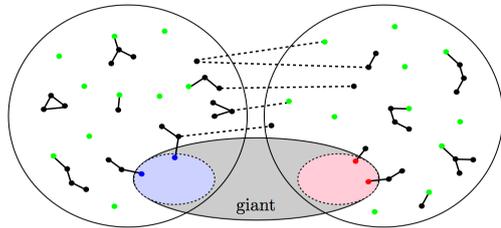


Fig. 2: Illustration of the topology of $\text{SBM}(n, k, a, b)$ for $k = 2$. A giant component covering the two communities takes place when $d = \frac{a+(k-1)b}{k} > 1$; a linear fraction of vertices belong to isolated trees (including isolate vertices). To estimate the size of the typical set: sample a bit uniformly at random in each isolated tree (green vertices) and propagate the bit according to the symmetric channel with flip probability $b/(a+(k-1)b)$ (plain edges do not flip whereas dashed edges flip).

We hence need to count the number of nodes T and edges M that belong to such trees in the SBM graph. This is done in a series of lemmas in our arxiv paper [6], and requires combinatorial estimates similar to those carried for the Erdős-Rényi case [16]. The fractions are shown to concentrate around

$$T/n \approx \frac{\tau}{d} \left(1 - \frac{\tau}{2}\right), \quad (5)$$

$$M/n \approx \frac{\tau^2}{2d}, \quad (6)$$

where τ is as in the theorem. This in turn gives a bound on the typical set size (see Theorem 2 below). With this bound, we can better estimate the probability of sampling a good clustering, reaching the tight bound at $b = 0$.

Theorem 2. Let $T_\delta(G)$ denote the typical set for G drawn under $\text{SBM}(n, k, a, b)$. Then, for any $\varepsilon > 0$,

$$\mathbb{P}\{|T_\delta(G)| < k^{(\psi-\varepsilon)n}\} = o(1),$$

where

$$\psi := \frac{\tau}{d} \left(1 - \frac{\tau}{2}\right) + \frac{\tau^2}{2d \ln k} H(\nu),$$

$$\nu := \left(\frac{a}{a+(k-1)b}, \frac{b}{a+(k-1)b}, \dots, \frac{b}{a+(k-1)b}\right)$$

and $H(\cdot)$ is the entropy in nats.

Proof sketch of Theorem 2: Let $G \sim \text{SBM}(n, k, a, b)$, and let T be the number of isolated trees in G , M the number of edges in those trees, and F the number edges in the planted trees of the largest connected component of G (i.e., the giant). We now build a typical assignment on these trees:

- Pick an arbitrary node in each isolated tree, denote these by $\{v_1, \dots, v_T\}$, and denote the set of edges contained in these trees by $\{E_1, \dots, E_M\}$;
- Assign the labels $U_1^T := (U_{v_1}, \dots, U_{v_T})$ uniformly at random in $[k]$. Then broadcast each of these labels in their corresponding trees by forwarding the labels on each edge with an independent k -ary symmetric channel of flip probability $\frac{b}{a+(k-1)b}$. This means that the variables Z_1, \dots, Z_M are drawn i.i.d. from the distribution ν as above on $\mathbb{F}_k := \{0, 1, \dots, k-1\}$, and that for each edge e in the trees, the input bit is forwarded by adding to it the Z_e variable modulo k ;
- Assign any other vertex (that is not contained in the trees) to their true community assignments. Define $Z_1^M := (Z_1, \dots, Z_M)$, and denote by $\hat{X}(U_1^T, Z_1^M)$ the previously defined assignment.

Note that the above gives the induced label-distribution on trees in $\text{SBM}(n, k, a, b)$, with possibly a global flip for the isolated trees. Thus, with high probability on G , as T and M grow (linearly) with n , this assignment is typical with high probability on U_1^T, Z_1^M :

$$\mathbb{P}_{U_1^T, Z_1^M} \{\hat{X}(U_1^T, Z_1^M) \in T_\delta(G)\} = 1 - o(1). \quad (7)$$

Denote by $A_{\varepsilon, M}(\nu)$ the ε -typical set for sequences of length M under the distribution η on $[k]$ (as defined in [17]). Define similarly $A_{\varepsilon, T}(\eta)$ for the uniform distribution η on k . By the AEP theorem, for any $\varepsilon > 0$,

$$\mathbb{P}\{U_1^T \in A_{\varepsilon, T}(\eta), Z_1^M \in A_{\varepsilon, M}(\nu)\} \rightarrow 1.$$

Therefore,

$$\begin{aligned} & \mathbb{P}_{U_1^T, Z_1^M} \{\hat{X}(U_1^T, Z_1^M) \in T_\delta(G)\} \\ & \leq \sum_{u_1^T \in A_{\varepsilon, T}(\eta), z_1^M \in A_{\varepsilon, M}(\nu)} \\ & 1(\hat{X}(u_1^T, z_1^M) \in T_\delta(G)) k^{-T} k^{-M(H(\nu)-\varepsilon)} + o(1). \end{aligned}$$

Since the right hand side counts a subset of the typical clusterings, we have with high probability on G ,

$$|T_\delta(G)| \geq (1 - o(1)) k^{T+M(H(\nu)-\varepsilon)}.$$

Further, from the topological lemmas derived in our arxiv

paper [6], for $\varepsilon > 0$, with high probability on G ,

$$\begin{aligned} T & \in \left[\frac{\tau}{d} \left(1 - \frac{\tau}{2}\right) - \varepsilon, \frac{\tau}{d} \left(1 - \frac{\tau}{2}\right) + \varepsilon \right], \\ M & \in \left[\frac{\tau^2}{2d} - \varepsilon, \frac{\tau^2}{2d} + \varepsilon \right]. \end{aligned}$$

The claims follows from algebraic manipulations. \blacksquare

ACKNOWLEDGEMENTS

This work is partly supported by NSF CAREER Award CCF-1552131 and ARO grant W911NF-16-1-0051.

REFERENCES

- [1] P. W. Holland, K. Laskey, and S. Leinhardt. Stochastic block-models: First steps. *Social Networks*, 5(2):109–137, 1983.
- [2] T.N. Bui, S. Chaudhuri, F.T. Leighton, and M. Sipser. Graph bisection algorithms with good average case behavior. *Combinatorica*, 7(2):171–191, 1987.
- [3] M.E. Dyer and A.M. Frieze. The solution of some random NP-hard problems in polynomial expected time. *Journal of Algorithms*, 10(4):451 – 489, 1989.
- [4] Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Struct. Algorithms*, 31(1):3–122, August 2007.
- [5] 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*, 84:066106, December 2011.
- [6] 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*, December 2015.
- [7] 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.
- [8] L. Massoulié. Community detection thresholds and the weak Ramanujan property. In *STOC 2014: 46th Annual Symposium on the Theory of Computing*, pages 1–10, New York, United States, June 2014.
- [9] E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *Available online at arXiv:1311.4115 [math.PR]*, January 2014.
- [10] E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. *Available online at arXiv:1202.1499 [math.PR]*, 2012.
- [11] A. Coja-oghlan. Graph partitioning via adaptive spectral techniques. *Comb. Probab. Comput.*, 19(2):227–284, March 2010.
- [12] J. Xu Y. Chen. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv:1402.1267*, February 2014.
- [13] J. Neeman and P. Netrapalli. Non-reconstructability in the stochastic block model. *Available at arXiv:1404.6304*, 2014.
- [14] A. Montanari. Finding one community in a sparse graph. *arXiv:1502.05680*, 2015.
- [15] J. Banks and C. Moore. Information-theoretic thresholds for community detection in sparse networks. *ArXiv e-prints*, January 2016.
- [16] P. Erdős and A Rényi. On the evolution of random graphs. In *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, pages 17–61, 1960.
- [17] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Interscience, New York, 1991.