EFFICIENT ALGORITHM FOR EXACT RECOVERY OF VERTEX VARIABLES FROM EDGE MEASUREMENTS

AFONSO S. BANDEIRA

What do community detection in complex networks, structure from motion in computer vision, and registration of multiple images have in common? It turns out that all these problems can be formulated as inverse problems on graphs. precisely, we can associate each data unit (a node label, photo, or rotated image) to a graph node, and each pairwise measurement to an edge connecting the two nodes whose associated data units are being compared. The problem can then be posed as estimating, for each node, an unknown variable (such as a community membership, or a rotation) from relative information between pairs of such variables. It is fruitful to think of node variables as taking values in a group $\mathcal{G}$ and the relative measurements as revealing information about group ratios $g_i \left(g_j\right)^{-1} \in \mathcal{G}$ (this is particularly natural when dealing with rotations but also helpful when thinking about community detection).

While there is a significant amount of literature proposing, and sometimes analysing, algorithmic approachs for some of these problems, a fundamental understanding of when these problems are solvable, either from an information theoretical or a computational average-case complexity viewpoint, is lacking. This paper [ABBS14] takes an important first step in this direction by treating a simple, yet crucial, instance of this framework — the setting where the node labels take values in $\{\pm 1\}$ (i.e. $\mathcal{G} \cong \{\pm 1\} \cong \mathbb{Z}_2$).

In this setting, the problem is best viewed as a form of community detection on an observation graph $G = (V, E)$. The unknown vertex labels can be represented as a binary vector x^V of cluster memberships and the edge measurements y^E are noisy indications of whether pairs of nodes belong to the same cluster. More precisely, given a noise level $\varepsilon < \frac{1}{2}$, for each edge (i, j), y_{ij}^E indicates wether i and j belong or not to the same commmunity, making an error with probability ϵ (independently for each edge). This is conviniently modeled by setting $y_{ij}^E = x_i \oplus x_j \oplus Z_{ij}$ where Z_{ij} are i.i.d. Bernoulli(ε) random variables and $\oplus$ represents the binary XOR. The goal is then to understand for which graphs G and values of ε can one expect to be able to recover x^V given y^E, and whether this inverse problem can be solved efficiently. It is worth noting that x^V can only be recovered up to a global flip (correponding to a cluster relabeling). Indeed, since only relative information is available, the measurements are invariant to relabeling of the clusters, for this reason recovering x^V is always meant modulo this transformation.

This model is tightly connected with the popular stochastic block model for two communities, where a random graph is drawn on vertices belonging to two (or more) communities from a distribution where edges are present independently and with probability p if between two vertices of the same community, and q otherwise. The objective here is gain to recover the community memberships. The main difference between this model and the one above is that, for the stochastic block model, every pair of nodes provides information (the non-existence of an edge is also itself information), while this is not the case in the model discussed here, where pairs of nodes not connected in G provide no information. For this reason this model has been refered to as the censored block model [AM13, HLM12]. Regardless of its differences, the techniques in this paper (both the information theoretical

Date: August 2015.

This is an earlier version of a manuscript that is to appear in IEEE Computer. A. S. Bandeira was supported by NSF grant DMS-1317308.

limits and the algorithmic part) have now been successfully adapted to the setting of the stochastic block model [ABH14].

One of the main contributions of the paper is to show that, if the observation graph is taken to be an Erdős-Rényi graph [1] with edge probability p (meaning that the average degree is roughly np), then exact recovery of x^V is possible, with high probabiltiy, if and only if

$$\alpha := np/\log(n) > 2/(1-2\varepsilon)^2 + o\left(2/(1-2\varepsilon)^2\right).$$

It is worth noting that, if $\alpha < 1$, the observation graph is known to contain isolated nodes with high probability, rendering recovery impossible even in a noiseless setting, as there would be nodes for which no information is available. The main tool behind establishing this result are concentration of measure inequalities.

On the algorithmic side, the quadratic nature of the maximum likelihood estimator for x^V can be leveraged to design a semidefinite programming relaxation based algorithm for this problem [Sin11]. Duality and estimates on spectral norms of certain random matrices [Tro12, Tro15] are used to show that this efficient algorithmic approach exactly recovers x^V, with high probability, at essentially a factor of 2 away from the information theoretical limit. Remarkably, this gap has since been closed by making use of sharper estimates on the spectrum of random matrices [Ban15, HWX15].

It is also worth pointing out that the techniques used to both understand the information theoretical limits and establish the efficacy of the semidefinite relaxation approach have also since been adapted to community detection with more than two communities [AS15, HWX15, ABKK15, PW15]. Also, while the problem considered here is to exactly recover x^V, recent work [CRV15, SKLZ15] analyzed the problem of partially recover x^V solving one of the open problems posed in the paper.

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.

[ABH14] E. Abbe, A. S. Bandeira, and G. Hall, *Exact recovery in the stochastic block model*, Available online at arXiv:1405.3267 [cs.SI] (2014).

[ABKK15] N. Agarwal, A. S. Bandeira, K. Koiliaris, and A. Kolla, *Multisection in the stochastic block model using semidefinite programming*, Available online at arXiv:1507.02323 [cs.DS] (2015).

[AM13] E. Abbe and A. Montanari, *Conditional random fields, planted constraint satisfaction and entropy concentration*, Proc. of RANDOM (Berkeley), August 2013, pp. 332–346.

[AS15] E. Abbe and C. Sandon, *Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms*, to appear in FOCS 2015, also available online at arXiv:1503.00609 [math.PR] (2015).

[Ban15] A. S. Bandeira, *Random Laplacian matrices and convex relaxations*, Available online at arXiv:1504.03987 [math.PR] (2015).

[CRV15] P. Chin, A. Rao, and V. Vu, *Stochastic block model and community detection in the sparse graphs: A spectral algorithm with optimal rate of recovery*, Available online at arXiv:1501.05021 [cs.DS] (2015).

[HLM12] S. Heimlicher, M. Lelarge, and L. Massoulié, *Community detection in the labelled stochastic block model*, arXiv:1209.2910 (2012).

[HWX15] B. Hajek, Y. Wu, and J. Xu, *Achieving exact cluster recovery threshold via semidefinite programming: Extensions*, Available online at arXiv:1502.07738 (2015).

[PW15] W. Perry and A. S. Wein, *A semidefinite program for unbalanced multisection in the stochastic block model*, Available online at arXiv:1507.05605 [cs.DS] (2015).

[Sin11] A. Singer, *Angular synchronization by eigenvectors and semidefinite programming*, Appl. Comput. Harmon. Anal. **30** (2011), no. 1, 20 – 36.

[SKLZ15] A. Saade, F. Krzakala, M. Lelarge, and L. Zdeborová, *Spectral detection in the censored block model*, Available online at arXiv:1502.00163 (2015).

[Tro12] J. A. Tropp, *User-friendly tail bounds for sums of random matrices*, Foundations of Computational Mathematics **12** (2012), no. 4, 389–434.

[Tro15] ———, *An introduction to matrix concentration inequalities*, Found. Trends Mach. Learning **8** (2015), no. 1–2, 1–230.

(Bandeira) Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Masachusetts 02142, USA (bandeira@mit.edu).

[1] An Erdős-Rényi graph with edge probability p is a random graph where each pair of nodes is connected by an edge, independently and with probability p