

Community Detection: Statistical Inference Models

Anupama Chowdhary
Principal Keen College,
Bikaner, Rajasthan, India
chowdharyanupama@gmail.com

Satya Prakash Sharma
Research scholar,
Bhagwant University Ajmer,
Rajasthan, India
sp1965_sharma@yahoo.co.in

Abstract:- Community detection in large networks through the methods based on the statistical inference model can identify the node community as well as find the interaction between the communities. Statistical inference based methods try to fit a generative model to the network data. This paper discusses the statistical inference methods which groups the communities on vertices or nodes.

I. INTRODUCTION:

Detecting clusters or communities in large real-world graphs such as large social networks in some automated manner is a problem of considerable interest. A community could be loosely described as a collection of vertices within a graph that are densely connected among themselves while being loosely connected to the rest of the graph [1]. Communities in a web graph for instance might correspond to sets of web sites dealing with related topics [2][3]. Community detection is a common area in graph data computations and data mining computations [4][5]. The first analysis of community structure was carried out by Weiss and Jacobson [6], who searched for work groups within a government agency. The authors studied the matrix of working relationships between members of the agency, which were identified by means of private interviews. Work groups were separated by removing the members working with people of different groups, which act as connectors between them. This idea of cutting the bridges between groups is at the basis of several modern algorithms of community detection.

II. GENERAL CONCEPTS:

Graph partitioning has been studied by several authors. Given a connected graph $G = (V, E)$ with edge weights $w_e \forall e \in E$, partition the node set V into n nonempty subsets so as to minimize the total weight of the edges with end points in two different subsets. This problem is known to be NP-hard in general [7].

Several authors including Barahona and Mahjoub [8] have studied the problem of partitioning a graph into at most two subsets. Kernighan and Lin [9] consider the problem where G is to be partitioned into at most k subsets and each subset has at most p nodes. This instance of the problem arises in VLSI layout design. They suggest an efficient heuristic for

this problem. Carlson and Nemhauser [10] consider the problem of partitioning a graph into at most k subsets with no restriction on the size of each subset. They formulate the problem as a quadratic program and suggest a method by which local minima may be obtained. Conforti et al. [11][12] have studied the equicut problem on a complete graph. Here $r = 2$ and we want at most $(\frac{1}{2} |V|)$ nodes in each subset. Grigtschel and Wakabayashi [12-14] have studied the problem when G is a complete graph and is to be partitioned into at most $|V|$ subsets. They have called it the clique partitioning problem. There is no restriction on the number of nodes in each subset. Given a graph $G = (V, E)$ an r -cut in G is a set of edges E' such that the graph $G' = (V, E - E')$ contains exactly r connected components. The problem of finding a minimum weight r -cut is known to be NP-hard in general [15]. Goldschmidt and Hochbaum [16] have shown the r -cut problem to be solvable in polynomial time on general graphs provided r is fixed and all edge weights are nonnegative. Basic concepts in graph theory can be found in Bondy and Murty [17].

Statistical inference is the method of deducing judgements about characteristics or parameters of the population. This inference is taken out by analyzing the sample drawn from the whole population. The conclusion of statistical inference is the proposition such as a point estimate, an interval estimate, cluster or classification of data points into groups etc. The proposition of our interest is cluster or classification of data points into groups.

Methods based on statistical inference attempt to fit a generative model to the network data, which not only serve as a description of the large-scale structure of the network, but also can be used to generalize the data and predict the occurrence of missing or spurious links in the network [18][19] and encodes the community structure. The overall

advantage of this approach compared to the alternatives is its more principled nature, and the capacity to inherently address issues of statistical significance. Most methods in the literature are based on the stochastic block model [20] as well as variants including mixed membership [21][22], degree-correction[23], and hierarchical structures[24]. Model selection can be performed using principled approaches such as minimum description length [25][26][20][21] or Bayesian model selection [22] and likelihood-ratio test [23]. Currently many algorithms exist to perform efficient inference of stochastic block models, including belief propagation [24][25] and agglomerative Monte Carlo [26].

Generative model is a model for generating all values for a phenomenon, both those that can be observed in the world and "target" variables that can only be computed from those observed. Some popular generative models are: Naive Bayes, Hidden Markov Models, Latent Dirichlet Allocation, Boltzmann Machines. Generative model can be used to perform prediction.

III. STATISTICAL INFERENCE MODEL FOR COMMUNITY DETECTION

Communities in a graph can be identified by

- grouping nodes or vertices (known as vertex community)
- grouping links or edges (known as edge community)

In this paper we will discuss vertex community based models.

Vertex community based models:

1. **Planted Partition Model:** This model is a generative model for random graphs. A graph $G = (V; E)$ generated according to this model has a hidden partition $V_1, \dots, V_k$ such that $V_1 \cup V_2 \cup \dots \cup V_k = V$ and $V_i \cap V_j = \emptyset$ for $i \neq j$. If a pair of nodes u and v both lie in some V_i , then, $\Pr[(u, v) \in E] = p$ otherwise $\Pr[(u, v) \in E] = r$. Thus, in the planted partition model, if u and v are two nodes in the same cluster, then their expected degrees are equal. In the planted partition problem, we are given a graph G generated by the planted partition model, and our goal is to find the hidden partition $V_1, \dots, V_k$ with high probability over graphs generated according to this model. There are various algorithms/methods developed on this model with complexity $O(N)$.

Algorithms/methods developed:

- Condon and Karp [27], present a simple, linear-time algorithm for the graph l -partition problem and analyze it on a random "planted l -partition" model. In this model, the n nodes of a graph are partitioned into l groups, each of size n/l ; two nodes in the same

group are connected by an edge with some probability p , and two nodes in different groups are connected by an edge with some probability $r < p$.

They show that if $p - r \geq n^{-\frac{1}{2} + \epsilon}$ for some constant ϵ , then the algorithm finds the optimal partition with probability $1 - \exp(-n^{\Theta(\epsilon)})$.

- Boppana [28] provide the algorithm based on computing eigenvalues and eigenvectors of matrices associated with the graph and ellipsoid algorithm, which is able to compute a lower bound for the bisection width and equals it for a class of random graphs. It is worth emphasizing that Boppana produces the optimal bisection. His algorithm with probability $1 - O(n^{-1})$ finds the minimum bisection for graphs in G_{nmb} with $\frac{1}{2}m - b \geq \frac{5}{2}\sqrt{mn \log n}$.
- Jerrum and Sorkin [29], they extend the study of the Metropolis algorithm to the problem of graph bisection, i.e., finding a partition of the vertex set of an undirected graph into two equal-sized sets so that the number of crossing edges is minimized. They states that with overwhelming probability the Metropolis process will converge to the unique minimum bisection in time about $O(n^2)$ provided $\Delta > 11/6$.
- McSherry [30] provided the planted partition problem. He provides a spectral algorithm to recover the planted partitions by using a projection of the nodes onto the top k eigenspace of the adjacency matrix. However, McSherry's work (McSherry, 2001) and those of his predecessors only address the case when all vertices in the same cluster have the same expected degree, and this method fails to recover the correct partition in graphs generated by the extended planted partition model when the degree distribution is too skewed [31].
- 2. **Newman's Mixture Model [32]:** This model is a stochastic one that parameterizes the probability of each possible configuration of group assignments and edges. They define θ_{ri} to be the probability that a (directed) link from a particular vertex in group r connects to vertex i . In the World Wide Web, for instance, θ_{ri} would represent the probability that a hyperlink from a web page in group r links to web page i . In effect θ_{ri} represents the "preferences" of vertices in group r about which other vertices they link to. In their approach it is these preferences that define the groups: a "group" is a set of vertices that all have similar patterns of connection to others. This method was further extended to undirected networks. He used expectation-maximization (EM) algorithm to implement this model. There are various algorithms/methods developed on this model

with complexity $O(KL)$, here K is number of communities and L is number of edges of the network.

Algorithms/methods developed:

- Yang, Sato, and Nakagawa [33] implemented EM-algorithm by Newman. They treated each vertex in the network as one party i.e. n -vertices become n -parties and assume that the network is connected. He performed his algorithm in two steps known as E-step and M-step. Time complexity is $O(RC(K + \log_K n))$, where R – rounds for E-step and M-step, C – in E-step each party i performs protocol 1 for all its children C times, K – maximum vertex degree, n – number of vertices.
- Vazquez [34][35] taking inspiration from mixture models, work on the Bayesian formulation of the problem of finding hypergraph communities.
- Mungan and Ramasco [36], their approach naturally allows for the identification of the key elements responsible for the grouping and their resilience to changes in the network. Given the generality of the assumptions underlying the statistical model, such nodes are likely to play special roles in the original system. They illustrate this point by analyzing several empirical networks for which further information about the properties of the nodes is available. The search and identification of stabilizing nodes constitutes thus a novel technique to characterize the relevance of nodes in complex networks.

3. Mixed Membership Model [37]: Mixed-membership models capture that

- Each group of data is built from the same components or, from a subset of the same components.
- How each group exhibits those components varies from group to group. Thus the model captures homogeneity and heterogeneity.

This involves the following (generic) generative process,

1. Draw components $\beta_k \sim f(\cdot | n)$.
2. For each group i :
 - (a) Draw proportions $\theta_i \sim \text{Dir}(\alpha)$.
 - (b) For each data point j within the group:
 - i. Draw a mixture assignment $z_{ij} \sim \text{Cat}(\theta_i)$.
 - ii. Draw the data point $x_{ij} \sim g(\cdot | \beta_{z_{ij}})$.

Algorithms/methods based on this method have complexity of $O(KN^2)$, where K – number of communities and N – number of nodes in network.

Algorithms/methods developed:

- Psorakis, et. al.[38][39] provide soft partitioning solutions, each node is associated with a membership distribution over communities, describing its degree of participation to each module and used Bayesian non-negative matrix factorization model.
- Parkkinen et al.[40] developed a model in which, the Bernoulli parameters of mixed membership stochastic block model [50] in the cells of the contingency table are replaced by a multinomial over all the cells. The model is able to generate multiple links for pairs of nodes, but on sparse graphs where the proportion of linked pairs p is small, the number of doubly linked pairs is on the order of p^2 , that is, vanishingly small, and the multinomial parameterization approximately corresponds to the Bernoulli parameterization.
- Blei, and Jordan[41], describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. They present approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation.
- Cohn and Chang [42] they describe a model of document citation that learns to identify hubs and authorities in a set of linked documents such as pages retrieved from the World Wide Web or papers retrieved from a research paper archive. Their model provides probabilistic estimates.
- Erosheva and Fienberg [43] developed general mixed-membership model that relies on four levels of assumptions: population, subject, latent variable and sampling scheme. The next assumption is whether the membership scores are treated as fixed or random in the model. Finally, the last level of assumptions specifies the number of distinct observed characteristics (attributes) and the number of replications for each characteristic.
- Yang et al.[44][45][46] present a probabilistic model for directed network community detection that aims to model both incoming links and outgoing links simultaneously and differentially. They introduce latent variables node productivity and node popularity to explicitly capture outgoing links and incoming links, respectively. We derive efficient EM algorithms for computing the maximum likelihood solutions to the proposed models.

- Nallapati et al., [47] they present two different models called the Pairwise-Link-LDA and the Link-PLSALDA models. The Pairwise-Link-LDA model combines the ideas of LDA [41] and Mixed Membership Block Stochastic Models [48] and allows modelling arbitrary link structure. Link-PLSALDA model combines the LDA and PLSA models.

4. **Mixed Membership Stochastic Block Model [49][50]:** Airoldi et al. introduced mixed membership stochastic block models, a novel class of latent variable models for relational data. They represent observed relational data as a graph $G = (N, Y)$, where $Y(p, q)$ maps pairs of nodes to values, that is, edge weights and consider binary matrices, where $Y(p, q) \in \{0, 1\}$. Graph $G = (N, Y)$ was drawn from the following procedure.

- For each node $p \in N$:
 - Draw a K dimensional mixed membership vector $\vec{\pi}_p \sim \text{Dirichlet}(\vec{\alpha})$.
- For each pair of nodes $(p, q) \in N \times N$:
 - Draw membership indicator for the initiator $\vec{z}_{p \rightarrow q} \sim \text{Multinomial}(\vec{\pi}_p)$.
 - Draw membership indicator for the receiver, $\vec{z}_{q \rightarrow p} \sim \text{Multinomial}(\vec{\pi}_q)$.
 - Sample the value of their interaction, $Y(p, q) \sim \text{Bernoulli}(\vec{z}_{p \rightarrow q} B \vec{z}_{q \rightarrow p})$.

Further they state, statistically each node is an admixture of group-specific interactions. The two sets of latent group indicators are denoted by $\{\vec{z}_{p \rightarrow q} : p, q \in N\} =: Z_{\rightarrow}$ and $\{\vec{z}_{p \leftarrow q} : p, q \in N\} =: Z_{\leftarrow}$. The joint probability of the data Y and the latent variables $\{\vec{\pi}_{1:N}, Z_{\rightarrow}, Z_{\leftarrow}\}$ was given by

$$P(Y, \vec{\pi}_{1:N}, Z_{\rightarrow}, Z_{\leftarrow} | \vec{\alpha}, B) = \prod_{p,q} P(Y(p, q) | \vec{z}_{p \rightarrow q}, \vec{z}_{p \leftarrow q}, B) P(\vec{z}_{p \rightarrow q} | \vec{\pi}_p) P(\vec{z}_{p \leftarrow q} | \vec{\pi}_q) \prod_p P(\vec{\pi}_p | \vec{\alpha})$$

The data can be thought of as a directed graph. This model provides exploratory tools for scientific analyses in applications where the observations can be represented as a collection of unipartite graphs. The nested variational inference algorithm is parallelizable and allows fast approximate inference on large graphs. Algorithms/methods based on this method have complexity of $O(KN^2)$, where K – number of communities and N – number of nodes in network.

5. **Degree-Corrected Stochastic Block Model [51][52]:** Karrer and Newman developed this model which generates networks with a given

number n of vertices and undirected edges divided among a given number K of communities. And color each community with a different color. They parameterized the model by a set of parameters θ_{iz} , which represent the propensity of vertex i to have edges of color z . Here, θ_{iz} is the expected number of edges of color z that lie between vertices i and j , they states that exact number being Poisson distributed about this mean value. The space required to store the parameter θ_{iz} require $O(nK)$ space. Given the true optimal values of θ_{iz} , the optimal values of $q_{iz}(z)$ were given by

$$q_{ij}(z) = \frac{\theta_{iz} \theta_{jz}}{\sum_z \theta_{iz} \theta_{jz}}$$

The optimal θ_{iz} was given by

$$\theta_{iz} = \frac{\sum_j A_{ij} q_{ij}(z)}{\sqrt{\sum_j A_{ij} q_{ij}(z)}}$$

The overall complexity of this method is $O(NK^2)$, where K – number of communities and N – number of nodes in network.

IV. CONCLUSIONS:

By analysing these models we can conclude that they can be used to

- Find overlapping and non-overlapping communities
- Find the relationship matrix between the community and the generalized community

These generation models are implemented via EM algorithm so they have high complexity. The networks have hierarchical structure but the existing statistical inference model cannot consider the network structure. If the data sets have nature of traditional community detection, then these methods can not produce appropriate results.

REFERENCES

- J. Bagrow and E. Bolt, "Local method for detecting communities". Physical Rev. E, Volume 72 No. 4, p 046108, 2015.
- D. Gibson, J. Kleinberg, and P. Raghavan, Inferring web communities from link topology. In Proceedings of the 9th ACM Conference on Hypertext and Hypermedia, Association of Computing Machinery, New York (1998).
- G. W. Flake, S. R. Lawrence, C. L. Giles, and F. M. Coetzee, Self-organization and identification of Web communities. IEEE Computer 35, 66(71) (2002)
- Shini Renjith, c Anjali, "Fitness function in genetic algorithm based information retrieval: A Survey", ICMIC 13, Dec. 2013, pp 80-86.
- Dhanaya Sudhakaran, Shini Renjith, "Phase based resource aware scheduler with job profiling for map reduce". TJLTET, Volume 6, issue 2, Nov. 2015, pp 92-96.

- [6] R.S. Weiss, E. Jacobson, A method for the analysis of the structure of complex organizations, *Am. Sociol. Rev.* 20 (1955) 661668.
- [7] M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, New York, 1979).
- [8] F. Barahona and A.R. Mahjoub, "On the cut polytope," *Mathematical Programming* 36 (1986) 157–173.
- [9] B.W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *Bell Systems Technical Journal* 49 (1970) 291–308.
- [10] R.C. Carlson and G.L. Nemhauser, "Scheduling to minimize interaction costs," *Operations Research* 14 (1966) 52–58.
- [11] M. Conforti, M.R. Rao and A. Sassano, "The Equipartition Polytope I," *Mathematical Programming* 49 (1990/91) 49–70.
- [12] M. Conforti, M.R. Rao and A. Sassano, "The Equipartition Polytope II," *Mathematical Programming* 49 (1990/91) 71–90.
- [13] M. Grötschel and Y. Wakabayashi, "Facets of the clique partitioning polytope," *Mathematical Programming* 47 (1990) 367–387.
- [14] M. Grötschel and Y. Wakabayashi "A cutting plane algorithm for a clustering problem," *Mathematical Programming* 45 (1989) 59–96. 42. M. Grötschel and Y. Wakabayashi "Composition of facets of the clique partitioning polytope," Report No. 18, Institut für Mathematik, Universität Augsburg (1987).
- [15] D.S. Johnson, C.H. Papadimitriou, P. Seymour and M. Yannakakis, "The complexity of multi way cuts," extended abstract (1983)
- [16] O. Goldschmidt and D.S. Hochbaum, "Polynomial algorithm for thek-cut problem," preprint, University of California (Berkeley, CA, 1987).
- [17] J.A. Bondy and U.S.R. Murty, *Graph Theory with Applications* (North-Holland, Amsterdam, 1976).
- [18] Guimerà, Roger; Marta Sales-Pardo (2009-12-29). "Missing and spurious interactions and the reconstruction of complex networks". *Proceedings of the National Academy of Sciences*. 106 (52): 22073–22078. PMC 2799723. PMID 20018705. doi:10.1073/pnas.0908366106. Retrieved 2011-11-09.
- [19] Clauset, Aaron; Cristopher Moore; M. E. J. Newman (2008-05-01). "Hierarchical structure and the prediction of missing links in networks". *Nature*. 453(7191): 98–101. ISSN 0028-0836. PMID 18451861. doi:10.1038/nature06830.
- [20] Holland, Paul W.; Kathryn Blackmond Laskey; Samuel Leinhardt (June 1983). "Stochastic blockmodels: First steps". *Social Networks*. 5 (2): 109–137. ISSN 0378-8733. doi:10.1016/0378-8733(83)90021-7. Retrieved 2011-08-26.
- [21] Airoldi, Edoardo M.; David M. Blei; Stephen E. Fienberg; Eric P. Xing (June 2008). "Mixed Membership Stochastic Blockmodels". *J. Mach. Learn. Res.* 9: 1981–2014. ISSN 1532-4435. Retrieved 2013-10-09.
- [22] Ball, Brian; Brian Karrer; M. E. J. Newman (2011). "Efficient and principled method for detecting communities in networks". *Physical Review E*. 84 (3): 036103. doi:10.1103/PhysRevE.84.036103. Retrieved 2011-12-08.
- [23] Karrer, Brian; M. E. J. Newman (2011-01-21). "Stochastic blockmodels and community structure in networks". *Physical Review E*. 83 (1): 016107. doi:10.1103/PhysRevE.83.016107. Retrieved 2011-11-08
- [24] Peixoto, Tiago P. (2014-03-24). "Hierarchical Block Structures and High-Resolution Model Selection in Large Networks". *Physical Review X*. 4 (1): 011047. doi:10.1103/PhysRevX.4.011047. Retrieved 2014-04-24.
- [25] P. Peixoto, T. (2017). "Bayesian stochastic blockmodeling". arXiv:1705.10225
- [26] P. Peixoto, T. (2013). "Parsimonious Module Inference in Large Networks". *Phys. Rev. Lett.* 110:148701. Bibcode:2013PhRvL.110n8701P. doi:10.1103/PhysRevLett.110.148701.
- [27] Condon, A., Karp, R.M., 2001. Algorithms for graph partitioning on the planted partition model. *Random Struct. Algorithms* 18 (2), 116–140
- [28] R.B. Boppana, Eigenvalues and Graph Bisection: An Average-Case Analysis, *IEEE Symposium on Foundations of Computer Science (FOCS)* (1987), 280-285.
- [29] M. Jerrum and G. B. Sorkin. Simulated annealing for graph bisection. In *Proceedings of the 34th Annual Symposium on Foundations of Computer Science*, pages 94–103, Washington, DC, 1993
- [30] F. McSherry. Spectral partitioning of random graphs. In *Proceedings of the 42nd Annual Symposium on Foundations of Computer Science*, pages 529–537, Las Vegas, Nevada, 2001
- [31] M. Mihail and C. H. Papadimitriou. On the eigenvalue power law. In *Proceedings of the 6th International Workshop on Randomization and Approximation Techniques*, pages 254–262, Cambridge, Massachusetts, 2002
- [32] M. E. J. Newman and E. A. Leicht, "Mixture models and exploratory analysis in networks", *Proceedings of the National Academy of Sciences*. 2007 Jun 5;104(23):9564-9.
- [33] Yang, Bin, Issei Sato, and Hiroshi Nakagawa. "Privacy-preserving EM algorithm for clustering on social network." *Advances in Knowledge Discovery and Data Mining* (2012): 542-553.
- [34] Vazquez, A., 2008. "Population stratification using a statistical model on hypergraphs". *Phys. Rev. E* 77 (6), 066106.
- [35] Vazquez, A., 2009. "Finding hypergraph communities: a Bayesian approach and variational solution". *J. Stat. Mech.: Theory Exp.*, 07006.
- [36] Mungan, M., Ramasco, J.J., 2010. Stability of maximum likelihood based clustering methods: exploring the backbone of classifications. *J. Stat. Mech.* 4, 04028.
- [37] Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- [38] Psorakis, Ioannis, Stephen Roberts, and Ben Sheldon. "Soft partitioning in networks via bayesian non-negative matrix factorization." *Adv Neural Inf Process Syst* (2010).

- [39] Psorakis, I., Roberts, S., Sheldon, B., 2010. Partitioning in networks via Bayesian Non-negative Matrix Factorization. In: The 24th Conference on NIPS
- [40] Parkkinen, J., Sinkkonen, J., Gyenge, A., et al., 2009. A block model suitable for sparse graphs. In: Proceeding of the 7th International Workshop on Mining and Learning with Graphs.
- [41] Blei, D.M., Ng, A.Y., Jordan, M.I., 2013. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993—1022.
- [42] Cohn, D., Chang, H., 2000. Learning to Probabilistically Identify Authoritative Documents *Citeseer.*, pp. 167—174
- [43] Erosheva, E., Fienberg, S., Lafferty, J., 2004. Mixed-membership models of scientific publications. *Proc. Natl. Acad. Sci. U. S. A.* 101, 5220—5223
- [44] Yang, T., Jin, R., Chi, Y., et al., 2009a. Combining link and content for community detection: a discriminative approach. *KDD*, 927—936.
- [45] Yang, T., Chi, Y., Zhu, S., et al., 2009b. A Bayesian framework for community detection integrating content and link. *BUAI*, 615—622.
- [46] 46. Yang, T., Chi, Y., Zhu, S., et al., 2010. Directed network community detection: a popularity and productivity link model. *SDM*, 742—753
- [47] Nallapati, R.M., Ahmed, A., Xing, E.P., et al., 2008. Joint latent topic models for text and citations. *ACM*, 542—550
- [48] M. Airoldi, D. Blei, E. Xing, and S. Fienberg. Mixed membership stochastic block models for relational data, with applications to protein-protein interactions. In *International Biometric Society-ENAR Annual Meetings*, 2006
- [49] Airoldi, E.M., Fienberg, S.E., et al., 2006. Mixed membership stochastic block models for relational data with application to protein—protein interactions. *Proc. Int. Biom. Soc.*
- [50] Airoldi, E.M., Blei, D.M., Fienberg, S.E., et al., 2008. Mixed membership stochastic block models. *J. Mach. Learn. Res.* 9, 1981—2014.
- [51] Karrer, B.B.B., Newman, M.E.J., 2011a. An efficient and principled method for detecting communities in networks. *Phys. Rev. E* 84 (3), 036103.
- [52] Karrer, B., Newman, M.E.J., 2011b. Stochastic block models and community structure in networks. *Phys. Rev. E* 83 (1), 016107.