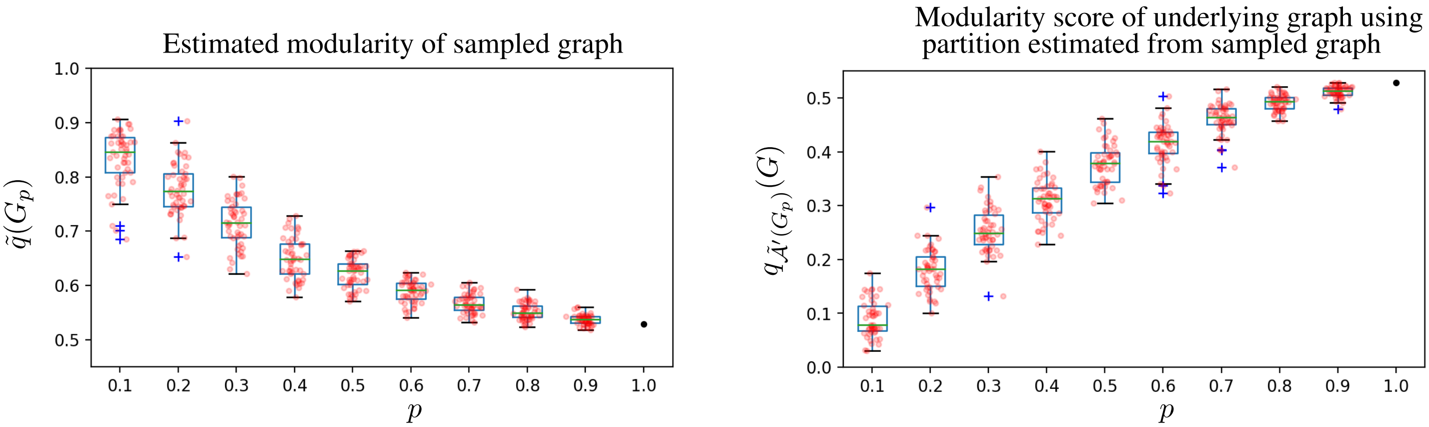
Partially observed graphs - when can we infer underlying community structure?

Colin McDiarmid[*] and Fiona Skerman[†]

Abstract

Suppose that there is an unknown underlying graph G on a large vertex set, and we can test only a proportion of the possible edges to check whether they are present in G. If G has high modularity, is the observed graph G' likely to have high modularity? We see that this is indeed the case under a mild condition, in a natural model where we test edges at random. We find that $q^*(G') \geq q^*(G) - \varepsilon$ with probability at least $1 - \varepsilon$, as long as the expected number edges in G' is large enough. Similarly, $q^*(G') \leq q^*(G) + \varepsilon$ with probability at least $1 - \varepsilon$, under the stronger condition that the expected average degree in G' is large enough. Further, under this stronger condition, finding a good partition for G' helps us to find a good partition for G.

We then consider the vertex sampling model for partially observing the underlying graph: we find that for dense underlying graphs we may estimate the modularity by sampling constantly many vertices and observing the corresponding induced subgraph, but this does not hold for underlying graphs with a subquadratic number of edges.


Estimated modularity of sampled graph

Modularity score of underlying graph using partition estimated from sampled graph

Figure 1: Simulation results. The dolphin social network [3] with 62 vertices and 159 edges was taken to be the underlying graph G. It is known that $q^*(G) = 0.529$ to three decimal places [2]. In the left part of the figure each red point corresponds to the estimated modularity $\tilde{q}(G_p)$ of an instance of the sampled graph G_p, i.e. the random graph formed by taking each edge in G to be present independently with probability p. For each edge probability $p = 0.1, 0.2, \ldots, 0.9$, the graph G_p was sampled 50 times. For each sampled graph G_p we took the maximum modularity score of the partitions output by over 200 runs of both the Louvain [1] and Leiden [4] algorithms. The noise in the x-axis is to allow one to see the points.
In the right part of the figure we examine, for each random instance of G_p, how well the modularity maximising partition of G_p performs as a partition on the underlying graph G. In detail, for each sampled graph G_p we plot the score $q_{\tilde{\mathcal{A}}'(G_p)}(G)$, where $\tilde{\mathcal{A}}(G_p)$ is the highest scoring partition on G_p found in 200 runs of Louvain and Leiden and $\tilde{\mathcal{A}}'(G_p)$ is the partition modified according to the method in our result.

References

[1] V. D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[2] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 20(2):172–188, 2008.

[3] D. Lusseau. The emergent properties of a dolphin social network. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(suppl_2):S186–S188, 2003.

[4] V. A. Traag, L. Waltman, and N. J. Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[*] Department of Statistics, University of Oxford. Email: *cmcd@stats.ox.ac.uk*

[†] Department of Mathematics, Uppsala University. Email: *fiona.skerman@math.uu.se*