$9^{th}$ International Conference on Computational Social Science $IC^2S^2$
July 17-20, 2023, Copenhagen, Denmark

# Entropy & fragmentation in socio-semantic networks

*Keywords: socio-semantic networks, block structures, metadata, entropy, partition dissensus*

## Extended Abstract

Online public spaces are often claimed to exhibit socio-semantic cohesiveness and fragmentation: gathering semantically similar actors into clusters, such spaces are – upon the influence of certain reinforcement dynamics – said to foster the emergence of 'echo chambers'. However, the literature that explores conversation networks in this context has been somewhat inconclusive in its findings, not only on the nature of socio-semantic fragmentation but even questioning its very existence. This calls for more comparative work to analyse if the emergence of socio-semantic fragmentation may in fact be context-dependent (e.g. on topic, platform, or interaction type), which in turn requires rigorous, statistically grounded methods to quantify such fragmentation in the first place. Generative models lend themselves particularly well in this context, since they can provide mathematical justification, help us deal with uncertainties connected to model choices and facilitate comparisons of different models and networks [Peel et al., 2022].

In this work, we introduce a method to quantify the existence and strength of *socio-semantic* fragmentation, based on measuring levels of 'assortative mixing' between *semantically* similar users in the *social* interaction network. For example, we assume high levels of fragmentation when actors who share political ideologies or other semantic similarities (represented by categorical node metadata) display a strong tendency to form social interactions with each other. Put simply, our measure indicates socio-semantic fragmentation if (structural) assortativity is the prominent structure on the network's mesoscale *and* the given semantic metadata correlates with this structure (i.e. semantic assortativity). Generative network models prevent well-known shortcomings of common assortativity measures such as modularity, which can be sensitive to network size and number of groups and which have been shown to find assortativity in random networks [Zhang and Peixoto, 2020]. Therefore, we use stochastic blockmodels (SBMs) and concepts from information theory to (a) identify whether assortative mixing between semantically similar actors is likely to be a key mechanism in the generative process of a given interaction network and, if so, to (b) measure the strength of this semantic assortativity.

Originally, the SBM was introduced as a generative model for networks with block structure [Holland et al., 1983] but it has since been repurposed as a key component of certain mesoscale structure detection methods, which use statistical inference to recover the most likely partition of a network given the generative process assumed by the SBM [Karrer and Newman, 2011, Peixoto, 2013]. SBMs are becoming increasingly popular partly due to their relatively general notion of node similarity which does not assume a particular type of mesoscale structure (such as assortativity) a priori. To identify the existence of socio-semantic fragmentation, our method uses a variant of the SBM tailored specifically to find assortative structure (aSBM) [Zhang and Peixoto, 2020]. It measures how well the aSBM – together with a partition according to semantic labels – explains the network at hand *compared to more general mesoscale partitions* and thus appraises whether assortativity between semantically similar actors is in fact the predominant structure in a network.

To formalise our measure, we thus need to compare how well the aSBM, given the semantic partition, does at explaining the network in relation to more general models. The description

length (DL) is a suitable way to compare models with potentially varying numbers of parameters, since – unlike pure network ensemble entropy – it prevents overfitting: it measures the amount of information needed to compress a network by combining an encoding of a model (and its parameters) with an encoding of the network *given* this model, and thus penalises overly complex models [Peixoto, 2013]. This serves our method in two major ways. It enables us to directly compare how well a set of metadata compresses the network compared to other sets of metadata and compared to partitions inferred by a metadata-agnostic SBM, even if partitions feature varying number of blocks. Additionally, the penalising part of the DL enables a comparison across different models for the same network, even when they differ in terms of their number of parameters.

In summary, our measure consists of two parts. We first establish the statistical significance of socio-semantic fragmentation in the network: motivated by Peel et al. [2017], we calculate the DL of the aSBM given the partition induced by semantic node metadata and compare it to the DL of distributions of a version of the same network with randomised metadata labels, to calculate a p-value. For networks that exhibit statistically significant fragmentation, we can calculate the second part of our measure to enable inter-network comparisons. For this purpose, we compute the 'compression ratio' between the description length of the network compressed by the optimal known (metadata-independent) partition and that of the network compressed by a given metadata partition, motivated by Kirkley [2022]. This measure allows us to compare entire collections of networks, for which we have the same sets of semantic metadata, in terms of the socio-semantic fragmentation strength.

We finally illustrate a use case of the method in the context of a meta-level comparison of topic-induced interaction networks on Twitter, to investigate the presence of topic groups in which conversations are more or less fragmented by socio-semantic clusters.

# References

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, June 1983. ISSN 03788733. doi: 10.1016/0378-8733(83)90021-7.

Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, January 2011. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.83.016107.

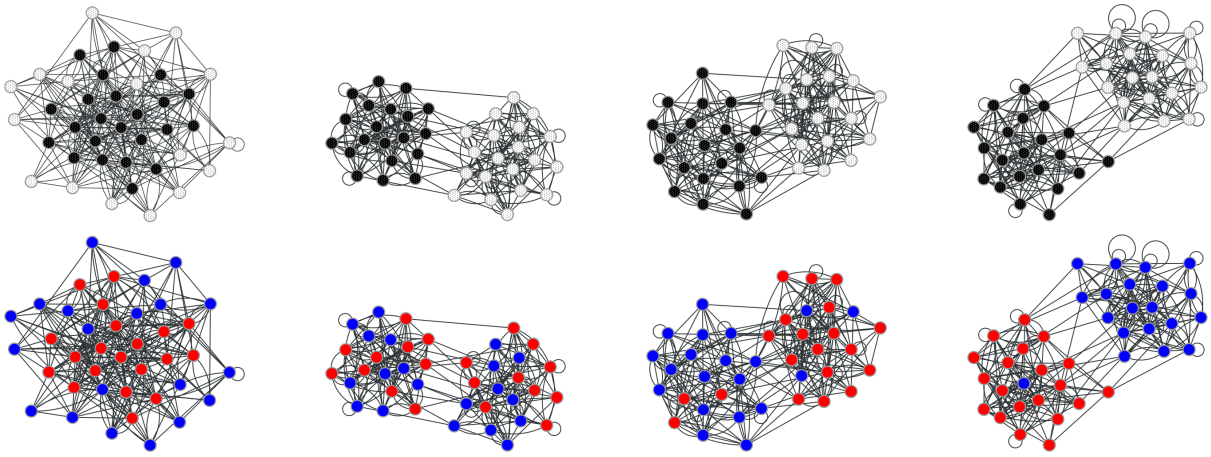
Alec Kirkley. Spatial regionalization based on optimal information compression. *Communications Physics*, 5(1):1–10, 2022. ISSN 2399-3650.

Leto Peel, Daniel B Larremore, and Aaron Clauset. The ground truth about metadata and community detection in networks. *Science Advances*, 3(5):e1602548, 2017.

Leto Peel, Tiago P. Peixoto, and Manlio De Domenico. Statistical inference links data and theory in network science. *Nature Communications*, 13(1):6794, 2022. ISSN 2041-1723.

Tiago P. Peixoto. Parsimonious module inference in large networks. *Physical review letters*, 110(14):148701, 2013.

Lizhi Zhang and Tiago P. Peixoto. Statistical inference of assortative community structures. *Physical Review Research*, 2(4):043271, November 2020. ISSN 2643-1564.



(a) Not significant (b) Not significant (c) Significant (d) Significant

Figure 1: Example networks that illustrate the socio-semantic fragmentation measure. For each network, nodes are coloured by the labels of the optimal (metadata-agnostic) partition (top) and the partition induced by semantic metadata labels (bottom). The networks in (a) and (b) do not exhibit significant socio-semantic fragmentation: in (a), the semantic labels correlate with the block structure but structural assortativity is not the prominent structure; in (b), structural assortativity is the prominent structure in the optimal partition of the network, but the semantic metadata does not correlate with it.
The networks in (c) and (d) both exhibit significant socio-semantic fragmentation, albeit to varying degrees: we expect the compression ratio (i.e. strength of fragmentation) of the network in (d) to be larger than that in (c), since the structural assortativity is more prominent and the metadata correlates more strongly with it.