Sampling Networks from Modular Compression of Network Flows

Christopher Blöcker[1,2], Jelena Smiljanić[3], Martin Rosvall[3], Ingo Scholtes[1,2]

[1]Chair of Machine Learning for Complex Networks, University of Würzburg, Germany [2]Data Analytics Group, Department of Informatics, University of Zürich, Switzerland [3]Integrated Science Lab, Department of Physics, Umeå University, Sweden

Summary

- Traditional models generate networks with given structural characteristics such as degree distribution, degree correlation, or community structure, but do not consider dynamical processes on networks
- Since dynamics are often superordinate to structure, we can learn about structure from dynamics
- We introduce a generative network model rooted in the modular compression of dynamic processes on networks as provided by the map equation and a related node similarity score

Background: The Map Equation and Map Equation Similarity

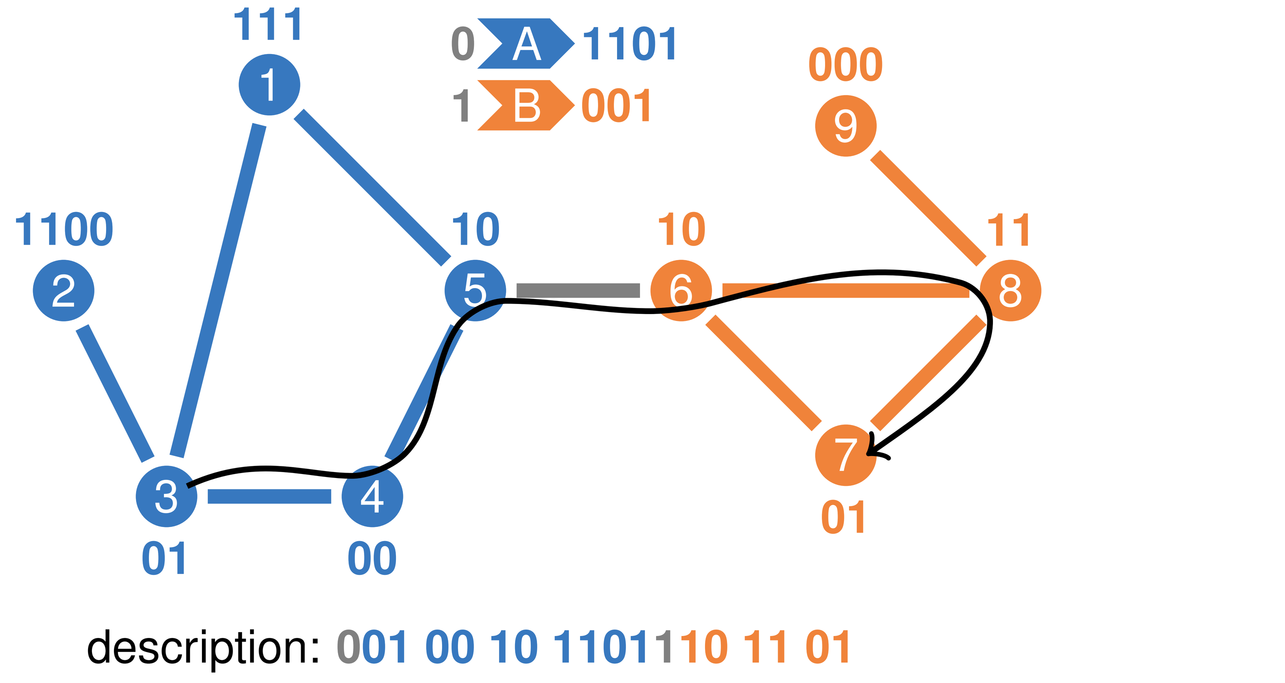

Figure 1: A network with two communities and a random walk trace.

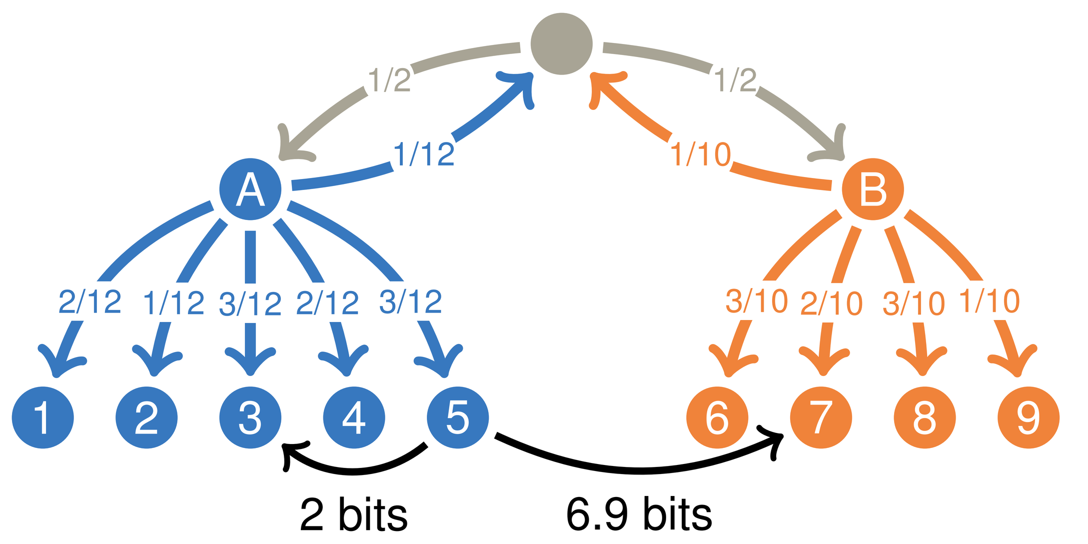

Figure 2: The network's community structure shown as a tree and annotated with the random walker's ergodic transition rates.

- The map equation (Rosvall & Bergstrom, 2008) is an information-theoretic objective function for community detection that relates a partition's goodness to the average per-step number of bits to describe random walks,

$$L(\mathsf{M}) = q\mathcal{H}(Q) + \sum_{\mathsf{m}\in\mathsf{M}} p_{\mathsf{m}}\mathcal{H}(P_{\mathsf{m}})$$

- A partition of the nodes into communities can be represented as a tree, annotated with transition rates
- For a partition M of the nodes into modules, MapSim quantifies the similarity between two nodes u and v, $\text{mapsim}(\mathsf{M}, u, v)$, as the rate at which a random walker at u visits v (Blöcker et al., 2022)
- We interpret $\text{mapsim}(\mathsf{M}, u, v)$ as a distance, $d_{uv} = -\log_2 \text{mapsim}(\mathsf{M}, u, v)$

We turn MapSim scores into link probabilities

- We use the softmax function to turn MapSim distances into link probabilities, $p_{uv} \propto k_u^{\text{out}} \cdot \frac{2^{-\beta d_{uv}}}{\sum_{v\neq u} 2^{-\beta d_{uv}}}$
- β is a temperature parameter that controls how peaked the resulting probability distribution is
- Expected out-degrees k_u^{out} are preserved, $E\left[k_u^{\text{out}}\right] = \sum_v k_u^{\text{out}} \cdot \frac{2^{-\beta d_{uv}}}{\sum_v 2^{-\beta d_{uv}}} = k_u^{\text{out}}$
- Expected in-degrees k_u^{in} are randomised, $E\left[k_u^{\text{in}}\right] = \sum_v k_v^{\text{out}} \cdot \frac{2^{-\beta d_{uv}}}{\sum_v 2^{-\beta d_{uv}}} \neq k_u^{\text{in}}$

Results

Table 1: Four networks with number of nodes $|V|$ and links $|E|$, average degree $\langle k\rangle$, number of communities $|\mathsf{M}|$, and mixing μ.

Network	Type	$\lvert V\rvert$	$\lvert E\rvert$	$\langle k\rangle$	$\lvert\mathsf{M}\rvert$	μ
Interactome	undirected	161	209	1.3	7	0.15
Highschool	directed	67	359	5.36	8	0.31
Anybeat	directed	8518	58799	6.9	542	0.53
arXiv citation HepPh	directed	12711	139981	11.01	270	0.38

- We sample networks based on MapSim link probabilities
- We use Infomap to detect communities in the original and sampled networks
- Community structure is preserved for large enough β, depending on network

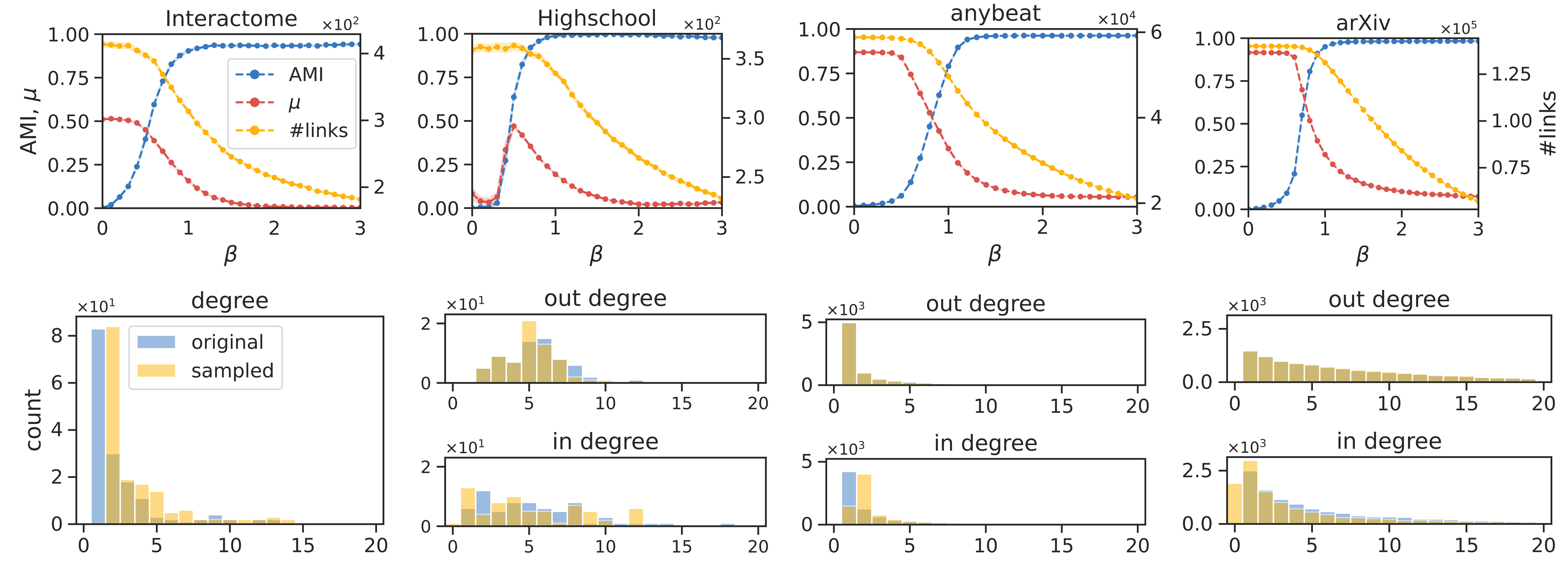

Figure 3: Each column shows the average AMI, average mixing μ, and the average number of links for 100 samples for each $\beta \in [0, 3]$, and the original and resulting degree distributions for $\beta = 1$.

Conclusion

- Our work connects generative and descriptive notions of community detection
- This approach can be used to generate benchmark networks based on dynamical processes for evaluating network analysis methods, adding to the scarce amount dynamics-based benchmark models
- In future work, we will investigate incorporating node metadata and higher-order dependencies into the network generation through map equation generalisations

Julius-Maximilians-UNIVERSITÄT WÜRZBURG

Universität Zürich[UZH]

UMEÅ UNIVERSITET