Flow Divergence: Comparing Network Partitions with Relative Entropy

Christopher Blöcker[1,2], Ingo Scholtes[2,1]

[1]Data Analytics Group, Department of Informatics, University of Zurich, Switzerland
[2]Chair of Machine Learning for Complex Networks, University of Würzburg, Germany

Summary

- Common partition-similarity measures ignore link patterns — instead they merely consider set overlaps with a reference partition, failing to distinguish between certain partitions
- To address this shortcoming, we combine the idea behind the Kullback-Leibler divergence with random walk descriptions and develop a link-aware partition-similarity measure, which we call *flow divergence*
- Applied to synthetic and real-world networks, we find that *flow divergene* distinguishes between network partitions that traditional measures consider to be equally good

The Problem

Common partition-similarity measures, such as the Jaccard index and adjuested mutual information (AMI), consider merely node labels but ignore link patterns. They judge partitions B, C, and D to be equally similar to the reference partition A.

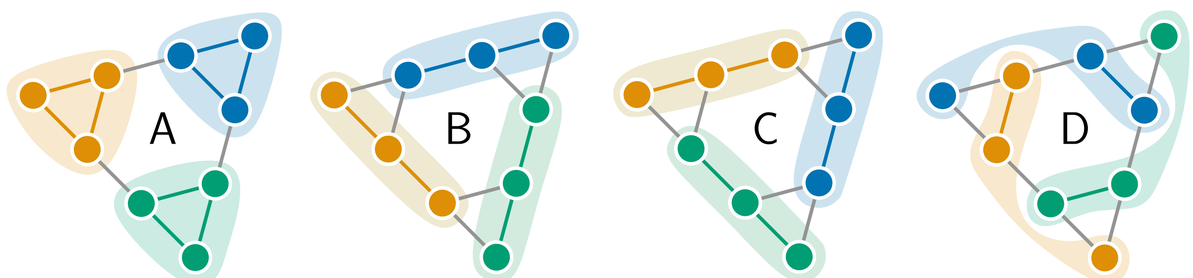

Figure 1: Four different partitions for the same network.

Our Solution: Flow Divergence

- We combine the ideas behind describing random walks and the Kullback-Leibler divergence to design a link-aware partition-similarity measure

$$D_{KL}(P \,||\, Q) = \sum_{x \in X} p_x \log_2 \frac{p_x}{q_x}$$

- *Flow divergence* quantifies the expected additional number of bits required to describe a random walk when using an estimate B of the network's "true" structure A

$$D_F(\mathrm{M}_a \,||\, \mathrm{M}_b) = \sum_{u \in V} p_u \sum_{v \in V} t_{uv}^{\mathrm{M}_a} \log_2 \frac{\mathrm{mapsim}(\mathrm{M}_a, u, v)}{\mathrm{mapsim}(\mathrm{M}_b, u, v)}$$

- mapsim is a node-similarity measure based on the map equation for community detection and quantifies how many bits are required to encode a random-walker step

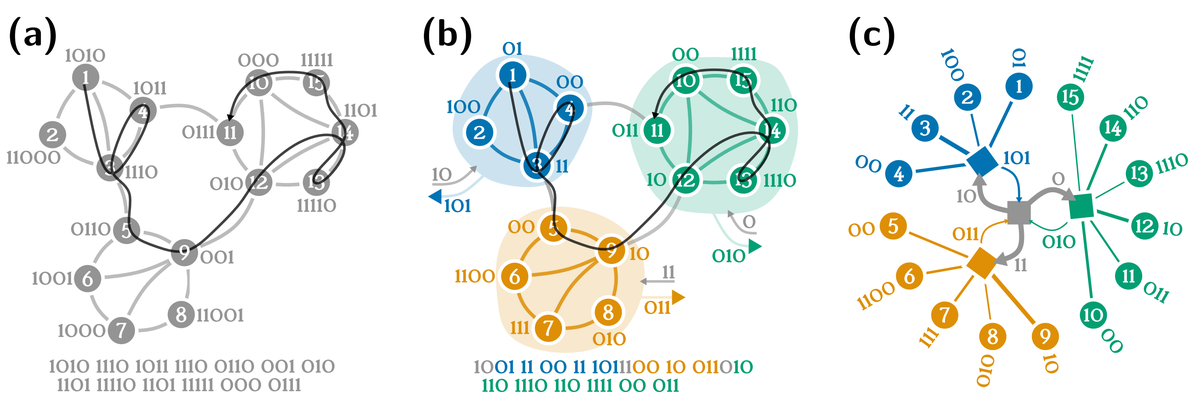

Figure 2: The coding principles behind the map equation where codewords are based on Huffman codes and the nodes' visit rates. **(a)** All nodes are in the same module and have unique codewords. **(b)** The network has three communities and nodes have unique codewords within modules. **(c)** The coding scheme from (b) shown as a radial tree.

Flow Divergence

- Based on random walks, *flow divergence* considers link patterns when comparing network partitions

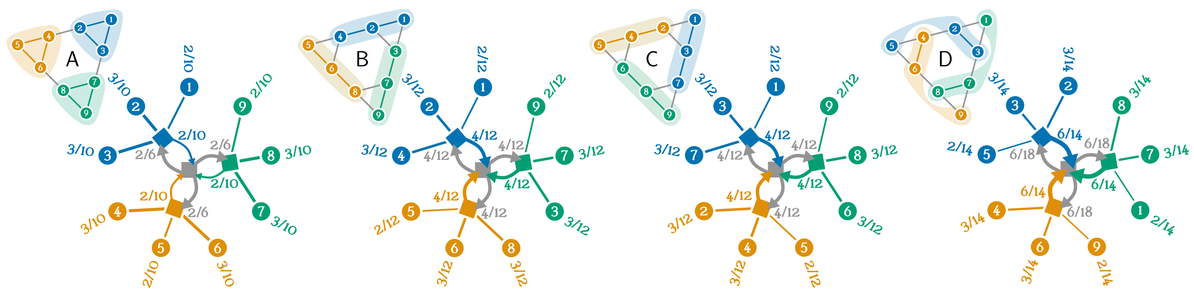

Figure 3: The four example partitions and their coding schemes, annotated with random-walker transition rates instead of codewords.

- Partition D is more similar to A than B and C because it requires fewer extra bits per step for describing a random walk

Table 1: Flow divergence in bits, rounded to two decimal places, between the partitions shown in Fig. 3.

	Other			
Reference	A	B	C	D
A	0	1.92	1.92	1.5
B	1.8	0	1.17	1.99
C	1.8	1.17	0	1.48
D	1.14	1.78	1.25	0

Application: Visualising Solution Landscapes

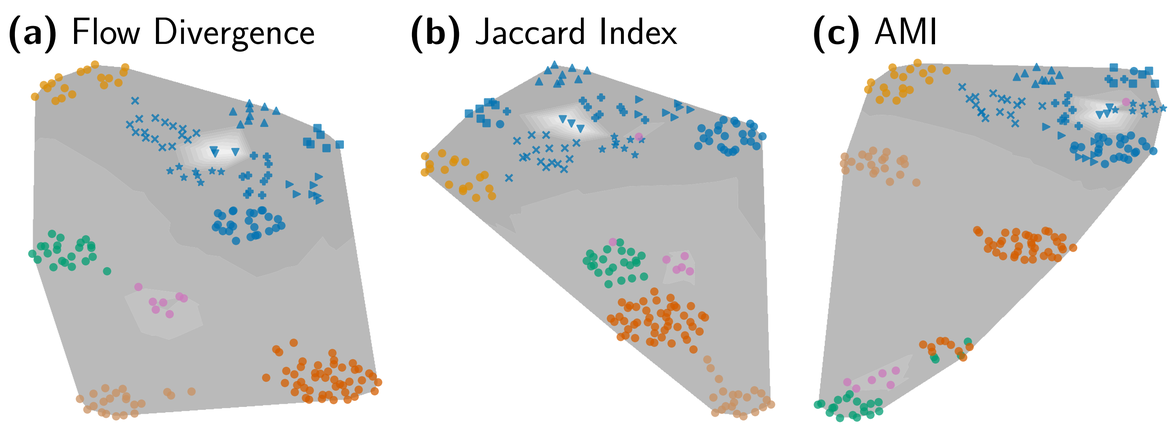

Figure 4: Three embeddings for 200 partitions of the Jazz network as identified with Infomap.

- We run Infomap with 200 different seeds to obtain 200 non-unique partitions for the Jazz network
- We compute the pairwise (dis-)similarities between all partitions with the Jaccard index, AMI, and flow divergence, and embed them with UMAP
- Different partition-similarity measures highlight different patterns in the solution landscape

Conclusion

- We designed *flow divergence* a partition-similarity measure based on the idea behind the Kullback-Leibler divergence and describing random walks
- *Flow divergence* considers link patterns when comparing network partitions and can distinguish between partitions where traditional measures fail

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Universität
Zürich[UZH]