Flow Divergence: Comparing Maps of Flows with Relative Entropy

NetSci 2024, Québec City, Canada

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

christopher.bloecker@uzh.ch

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

too long; didn't listen

- Networks can be partitioned in many different ways
  $\rightarrow$ we often compare partitions
- Common partition-similarity scores ignore link patterns
  $\rightarrow$ Jaccard index, variants of mutual information, ...
- We propose *flow divergence*
  - link-aware partition-similarity score
  - description of random walks + principle behind Kullback-Leibler divergence

Comparing Partitions

Comparing Partitions

Jaccard index $\quad J(a, b) = \frac{|a \cap b|}{|a \cup b|} \qquad J(A, B) = \sum_{a \in A} \frac{|a|}{n} \max_{b \in B} J(a, b)$

Reference: A

B $\qquad J(A, B) = \frac{1}{2}$

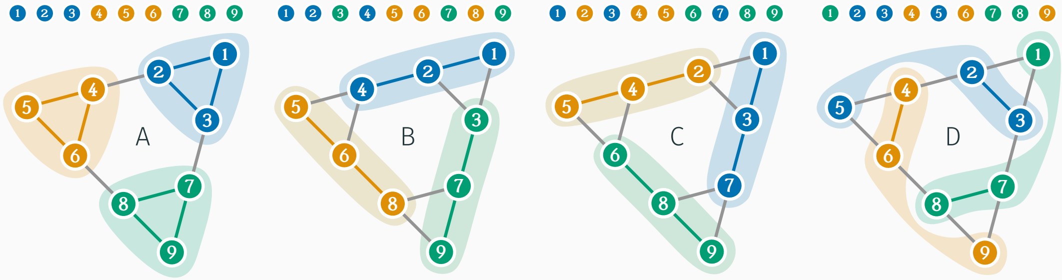
C $\qquad J(A, C) = \frac{1}{2}$

D $\qquad J(A, D) = \frac{1}{2}$

Similar for mutual information etc.

Comparing *Network* Partitions

Comparing Network Partitions



$$J\,(\mathrm{A},\mathrm{B}) = J\,(\mathrm{A},\mathrm{C}) = J\,(\mathrm{A},\mathrm{D}) = \frac{1}{2}$$

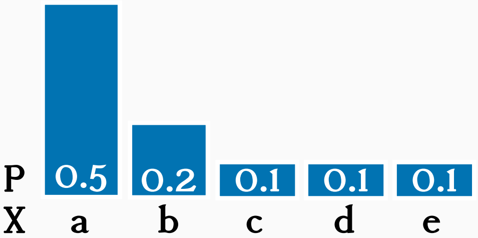
But what about the links?

Idea:
KL divergence + random walks

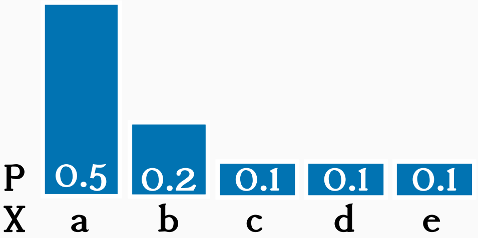
Kullback-Leibler divergence

X	**a**	**b**	**c**	**d**	**e**

Kullback-Leibler divergence

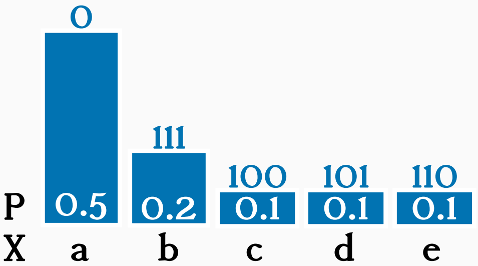


Kullback-Leibler divergence



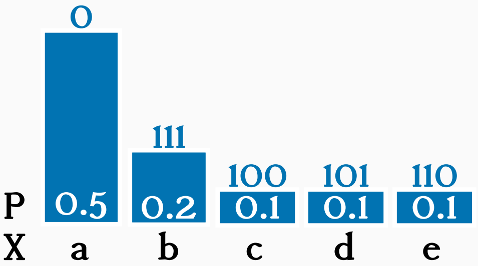
a a c d b b a b e c

Kullback-Leibler divergence



a a c d b b a b e c

Kullback-Leibler divergence

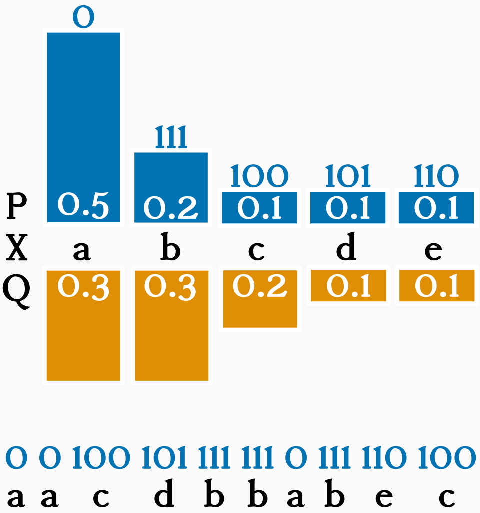


$$H(P) = -\sum_{x \in X} p_x \log_2 p_x$$

$$H(P) \approx 1.96 \text{ bits}$$

0 0 100 101 111 111 0 111 110 100

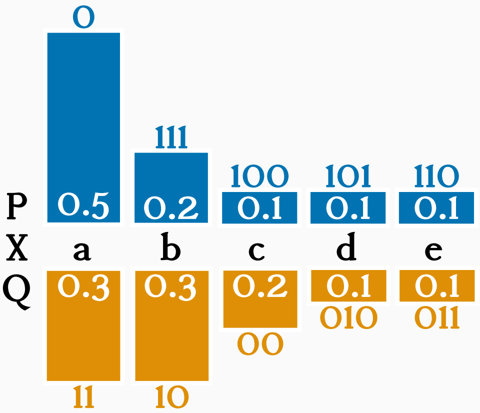
a a c d b b a b e c



$$H(P) = -\sum_{x \in X} p_x \log_2 p_x$$

$$H(P) \approx 1.96 \text{ bits}$$

Kullback-Leibler divergence



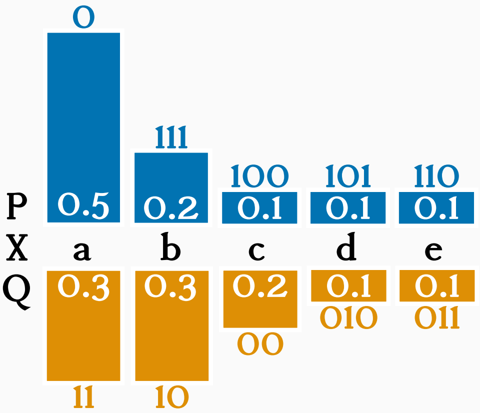
0 0 100 101 111 111 0 111 110 100

a a c d b b a b e c

$$H(P) = -\sum_{x \in X} p_x \log_2 p_x$$

$$H(P) \approx 1.96 \text{ bits}$$

Kullback-Leibler divergence



0 0 100 101 111 111 0 111 110 100

a a c d b b a b e c

11 11 00 010 10 10 11 10 011 00

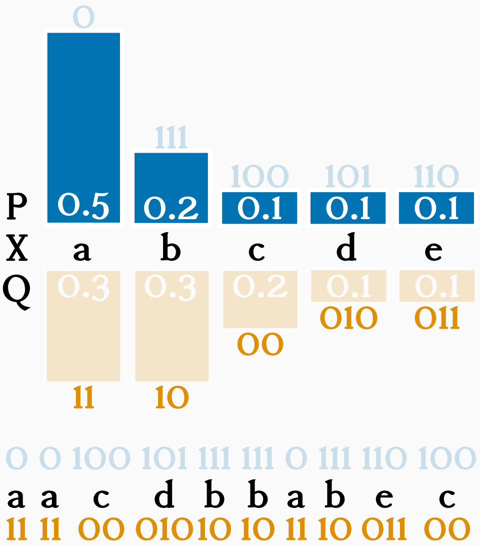
$$H(P) = -\sum_{x \in X} p_x \log_2 p_x$$

$$H(P) \approx 1.96 \text{ bits}$$

$$H(Q) \approx 2.17 \text{ bits}$$

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

Kullback-Leibler divergence



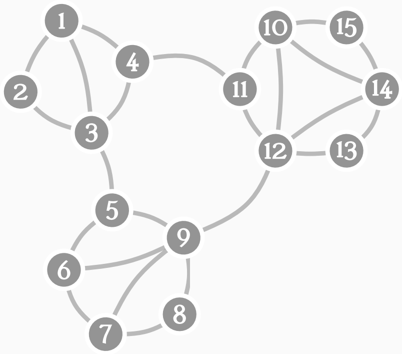
$$H(P) = -\sum_{x \in X} p_x \log_2 p_x$$

$$H(P) \approx 1.96 \text{ bits}$$

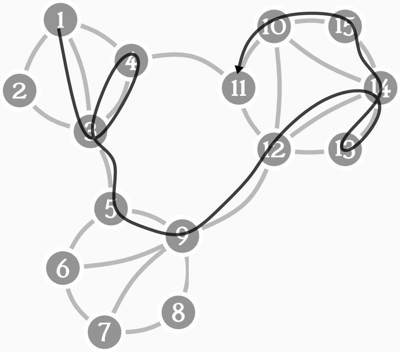
$$H(Q) \approx 2.17 \text{ bits}$$

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

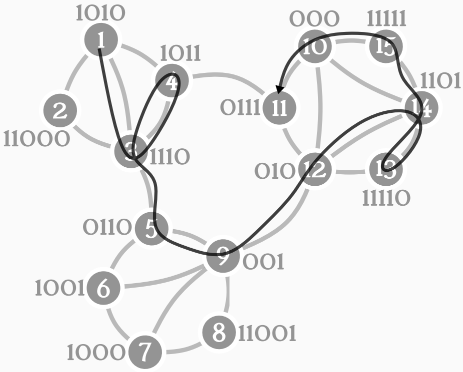
$$D_{KL}(P \,||\, Q) \approx 0.15 \text{ bits}$$

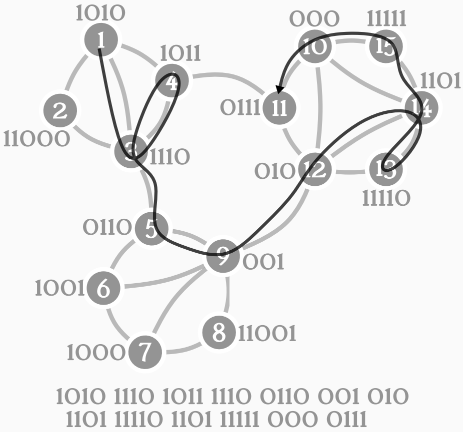

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Random Walks



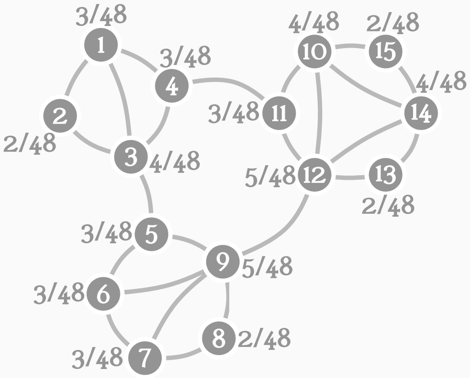
Random Walks





Description length

$$H(P) = -\sum_{v \in V} p_v \log_2 p_v \text{ bits} \tag{1}$$



Description length

$$H(P) = -\sum_{v \in V} p_v \log_2 p_v \text{ bits} \tag{1}$$

Visit rates

$$p_v = \sum_{u \in V} p_u t_{uv} \qquad t_{uv} = \frac{w_{uv}}{\sum_{v \in V} w_{uv}} \tag{2}$$

Combine (1)+(2)

$$H(P) = -\sum_{u \in V} p_u \sum_{v \in V} t_{uv} \log_2 p_v$$

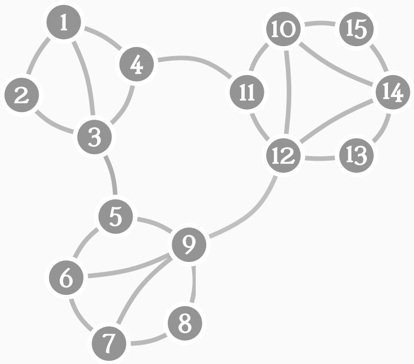
But we are interested in communities...

$$H(M) = -\sum_{u \in V} p_u \sum_{v \in V} t_{uv} \log_2 s(M, u, v)$$

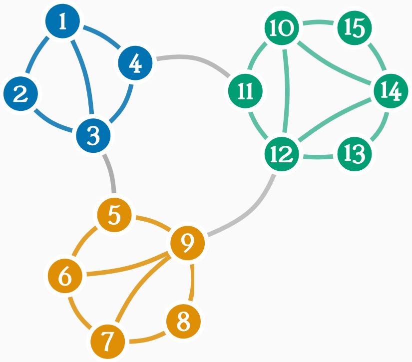
The Map Equation

$\longrightarrow$ Rosvall and Bergstrom; PNAS 2008

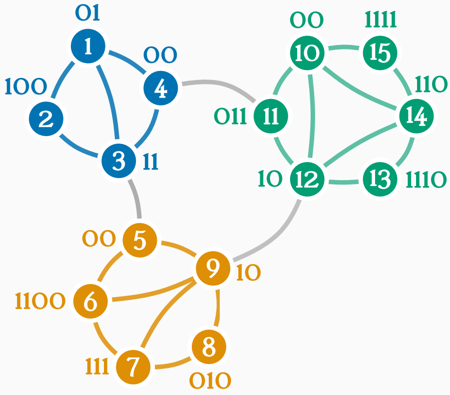
The Map Equation

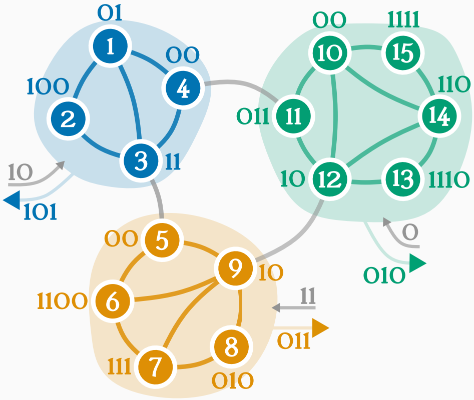


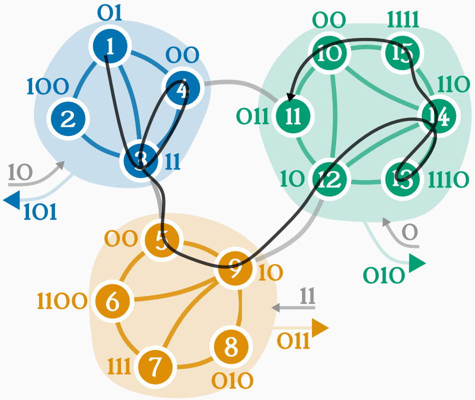
The Map Equation

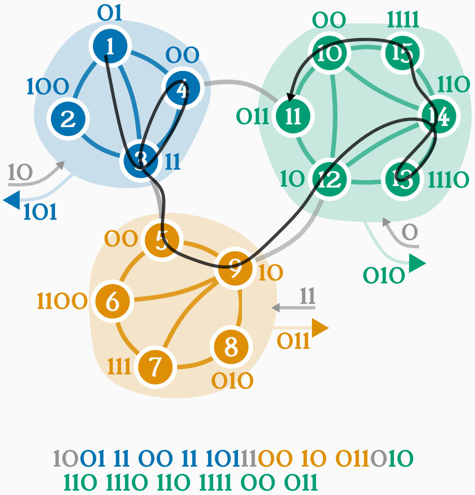


The Map Equation




01
00
100
1
4
2
3
11
10
101
00
1111
10
15
110
011
11
14
10
12
13
1110
0
010
00
5
9
10
1100
6
11
011
111
7
8
010

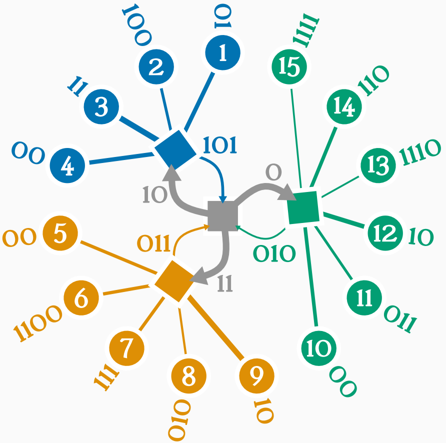

01
00
100
11
10
101
00
1111
110
011
10
1110
0
010
00
10
1100
11
011
111
010
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15



$$L(M) = qH(Q) + \sum_{m \in M} p_m H(P_m) \rightarrow \min$$

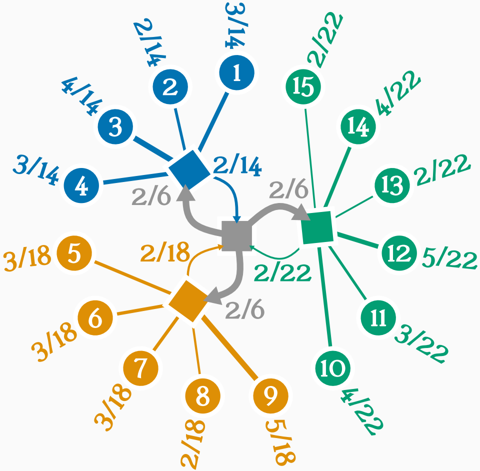
$$= -\sum_{u \in V} p_u \sum_{v \in V} t_{uv} \log_2 \text{mapsim}(M, u, v)$$

The Map Equation



1001 11 00 11 1011100 10 011010
110 1110 110 1111 00 011

$$
\begin{aligned}
L(\mathsf{M}) &= qH(Q) + \sum_{\mathsf{m} \in \mathsf{M}} p_{\mathsf{m}} H(P_{\mathsf{m}}) \\
&= -\sum_{u \in V} p_u \sum_{v \in V} t_{uv} \log_2 \operatorname{mapsim}(\mathsf{M}, u, v)
\end{aligned}
$$



$$L(\mathsf{M}) = qH(\mathsf{Q}) + \sum_{\mathsf{m}\in\mathsf{M}} p_{\mathsf{m}} H(P_{\mathsf{m}})$$
$$= -\sum_{u\in V} p_u \sum_{v\in V} t_{uv} \log_2 \mathbf{mapsim}(\mathsf{M}, u, v)$$

$$\mathbf{mapsim}(\mathsf{M}, u, v) = \begin{cases} \dfrac{p_v}{p_{\mathsf{m}_v}} & \text{if } \mathsf{m}_u = \mathsf{m}_v, \\[2ex] \dfrac{\mathsf{m}_{u,\text{exit}}}{p_{\mathsf{m}_u}} \cdot \dfrac{q_{\mathsf{m}_v}}{q} \cdot \dfrac{p_v}{p_{\mathsf{m}_v}} & \text{if } \mathsf{m}_u \neq \mathsf{m}_v, \end{cases}$$

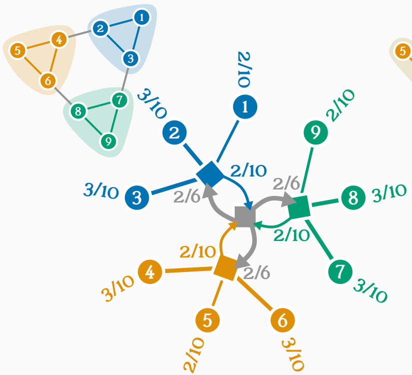
→ Blöcker et al.; PMLR 2022

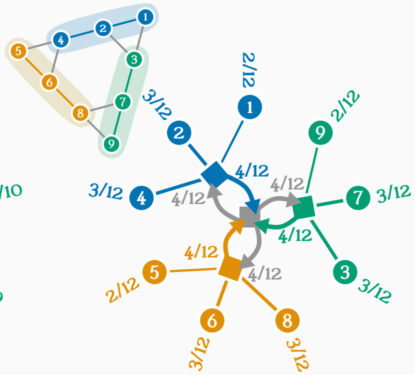
Flow Divergence

Flow Divergence

Idea

Measure how many extra bits we need to describe a random walk using an “estimate” M_b of the networks “true” partition M_a.





Flow Divergence

$$D_{KL}(P \,||\, Q) = \sum_{x \in X} p_x \log_2 \frac{p_x}{q_x} \qquad L(\mathsf{M}) = -\sum_{u \in V} p_u \sum_{v \in V} t_{uv} \log_2 \text{mapsim}(\mathsf{M}, u, v)$$

$$D(\mathsf{M}_a \,||\, \mathsf{M}_b) = \sum_{u \in V} p_u \sum_{v \in V} t_{uv} \log_2 \frac{\text{mapsim}(\mathsf{M}_a, u, v)}{\text{mapsim}(\mathsf{M}_b, u, v)} \quad ?$$

Not quite ... $D(\mathsf{M}_a \,||\, \mathsf{M}_b) = L(\mathsf{M}_a) - L(\mathsf{M}_b)$

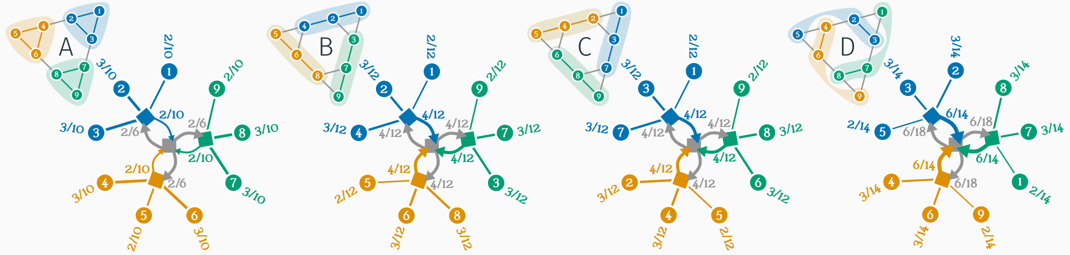
$$D_F(\mathsf{M}_a \,||\, \mathsf{M}_b) = \sum_{u \in V} p_u \sum_{v \in V} t_{uv}^{\mathsf{M}_a} \log_2 \frac{\text{mapsim}(\mathsf{M}_a, u, v)}{\text{mapsim}(\mathsf{M}_b, u, v)}$$

$$t_{uv}^{\mathsf{M}} = \frac{\text{mapsim}(\mathsf{M}, u, v)}{\sum_{v(\neq u)} \text{mapsim}(\mathsf{M}, u, v)}$$

Naïvely, the complexity is $\mathcal{O}(n^2)$ because we consider all pairs of nodes
$\rightarrow$ We're working on computing flow divergence more efficiently

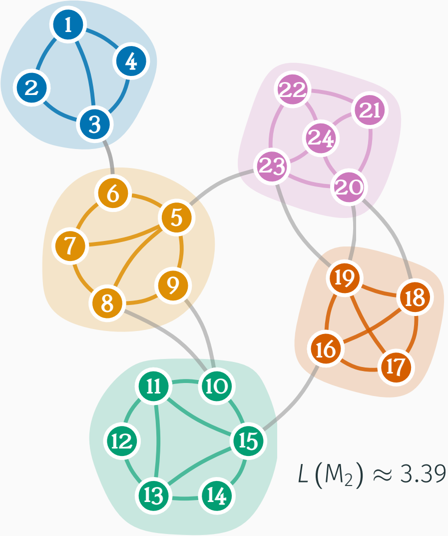
Flow Divergence

$L\,(\mathrm{A}) \approx 2.86$ bits $\qquad L\,(\mathrm{B}) \approx 3.73$ bits $\qquad L\,(\mathrm{C}) \approx 3.73$ bits $\qquad L\,(\mathrm{D}) \approx 4.47$ bits

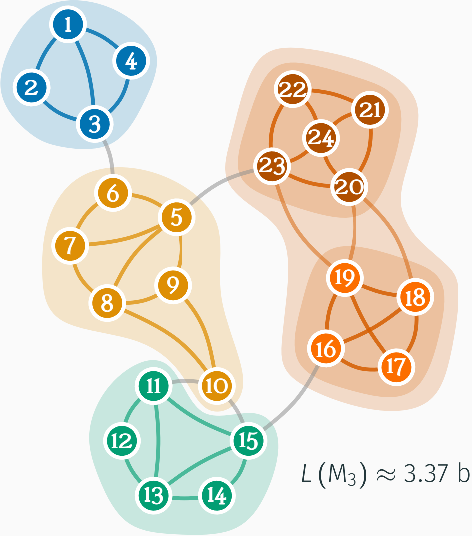


	Other			
Reference	A	B	C	D
A	0	1.92	1.92	1.5
B	1.8	0	1.17	1.99
C	1.8	1.17	0	1.48
D	1.14	1.78	1.25	0

Flow Divergence

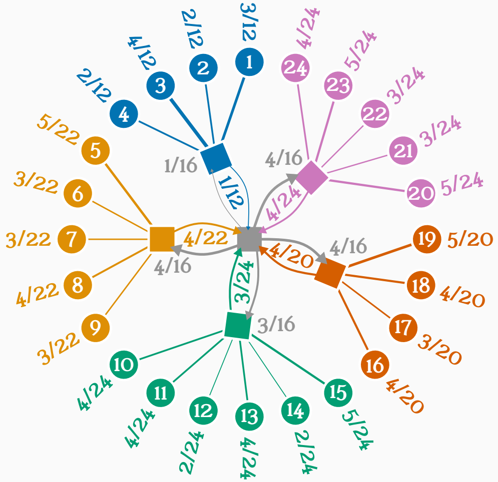


$L(\mathsf{M}_2) \approx 3.39$ bits

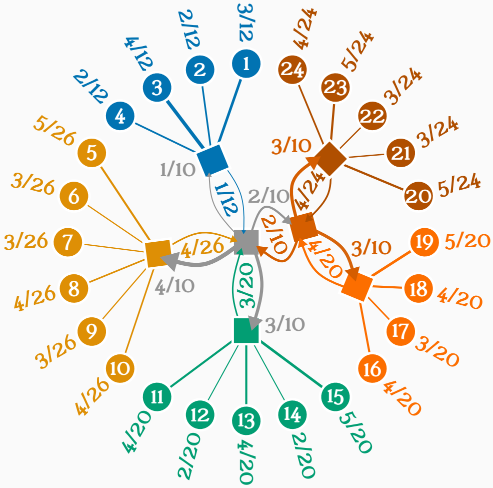


$L(\mathsf{M}_3) \approx 3.37$ bits

Flow Divergence



$D_F\,(M_2\,||\,M_3) \approx 0.11$ bits



$D_F\,(M_3\,||\,M_2) \approx 0.01$ bits

Conclusion

Conclusion

Motivation

Communities are based on link patterns
but common partition-similarity scores ignore links.

Our Solution

We combine the ideas behind describing random walks and the KL-divergence to develop a link-aware partition-similarity score.

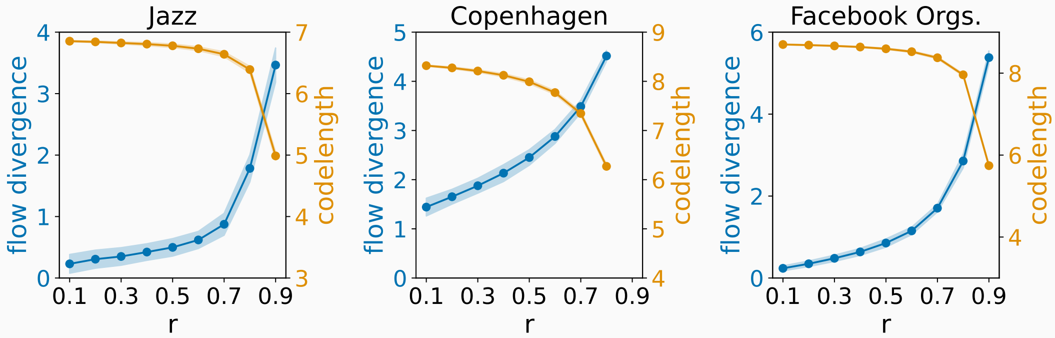
- asymmetric
- does not require matching of communities between partitions
- naturally compares hierarchical partitions

Thank you for your attention!

Check out our preprint: arXiv:2401.09052

Appendix

Flow Divergence



Flow Divergence

(a) Flow Divergence (b) Jaccard Index (c) Adjusted Mutual Information