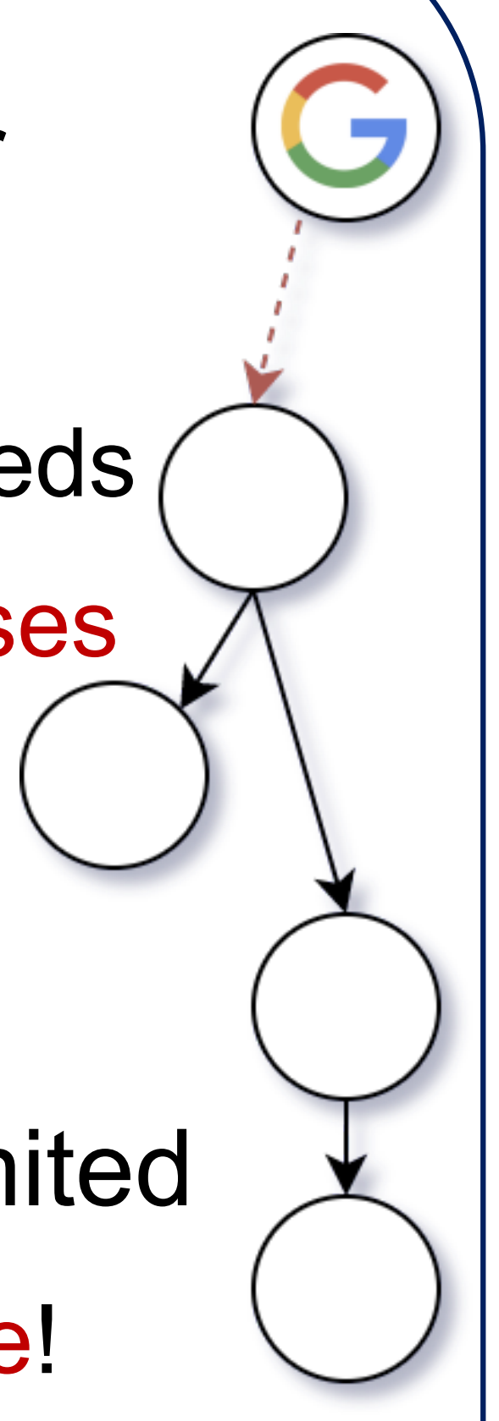
Wikipedia Reader Navigation: When Synthetic Data Is Enough

Akhil Arora, Martin Gerlach, Tiziano Piccardi, Alberto García Durán, and Robert West
Data Science Laboratory (DLAB), EPFL

Wikipedia reader navigation

"Information-rich" traces of reader behavior

- Understand and serve readers' needs
- Identify and mitigate structural biases
- Address knowledge gaps
- Organize articles into a curriculum



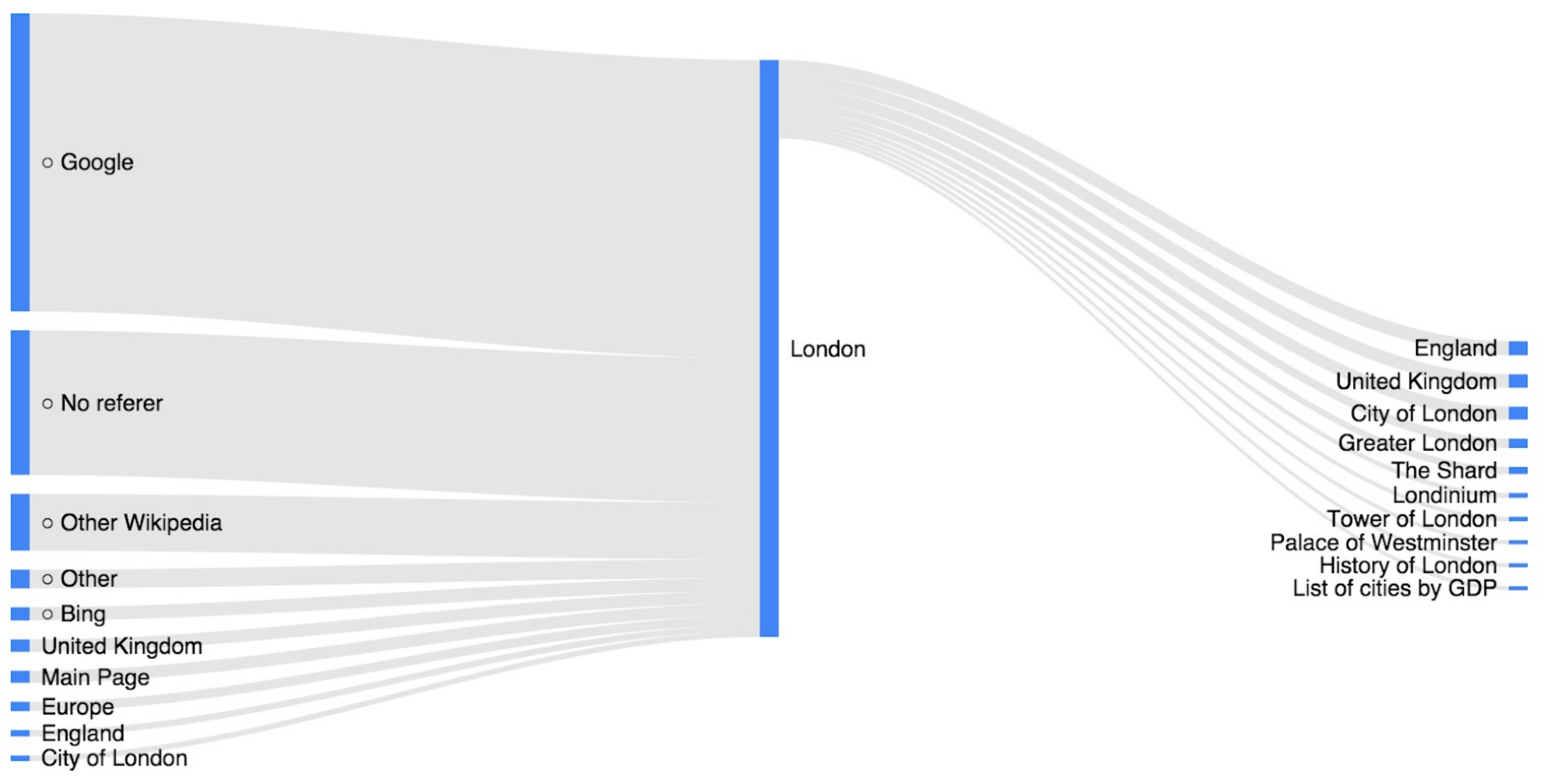
Studies on navigation data are limited

- Real traces are usually kept private!

What is Wikipedia Clickstream?

Public data consisting of

- Counts of (referrer, resource) pairs extracted from (private) server logs
- 1-hop neighborhood of each page
- Omits pairs occurring < 10 times



Only captures first order navigation behavior

- Next page visit depends only on the current page

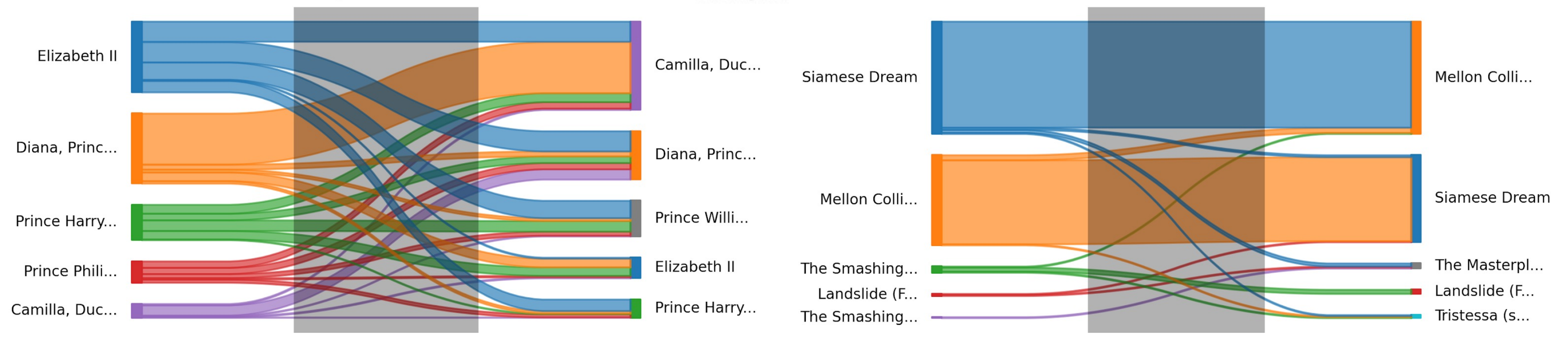
Key research questions

- How different are real trajectories from synthetic ones generated via Clickstream?
- How well can we approximate reader navigation via Wikipedia clickstream?

Dataset	Type	Main Characteristics
Logs	Real	Human navigation on Wikipedia.
Clickstream-Priv	Synthetic	Markov-1, biased random walks using private Clickstream.
Clickstream-Pub	Synthetic	Markov-1, biased random walks using public Clickstream.
Clickstream-Pub (I)	Synthetic	Markov-1, biased random walks using public Clickstream, with a different intrinsic stopping criterion [54].
Graph	Synthetic	Markov-1, unbiased random walks on Wikipedia hyperlink graph.

Mixing of Flows

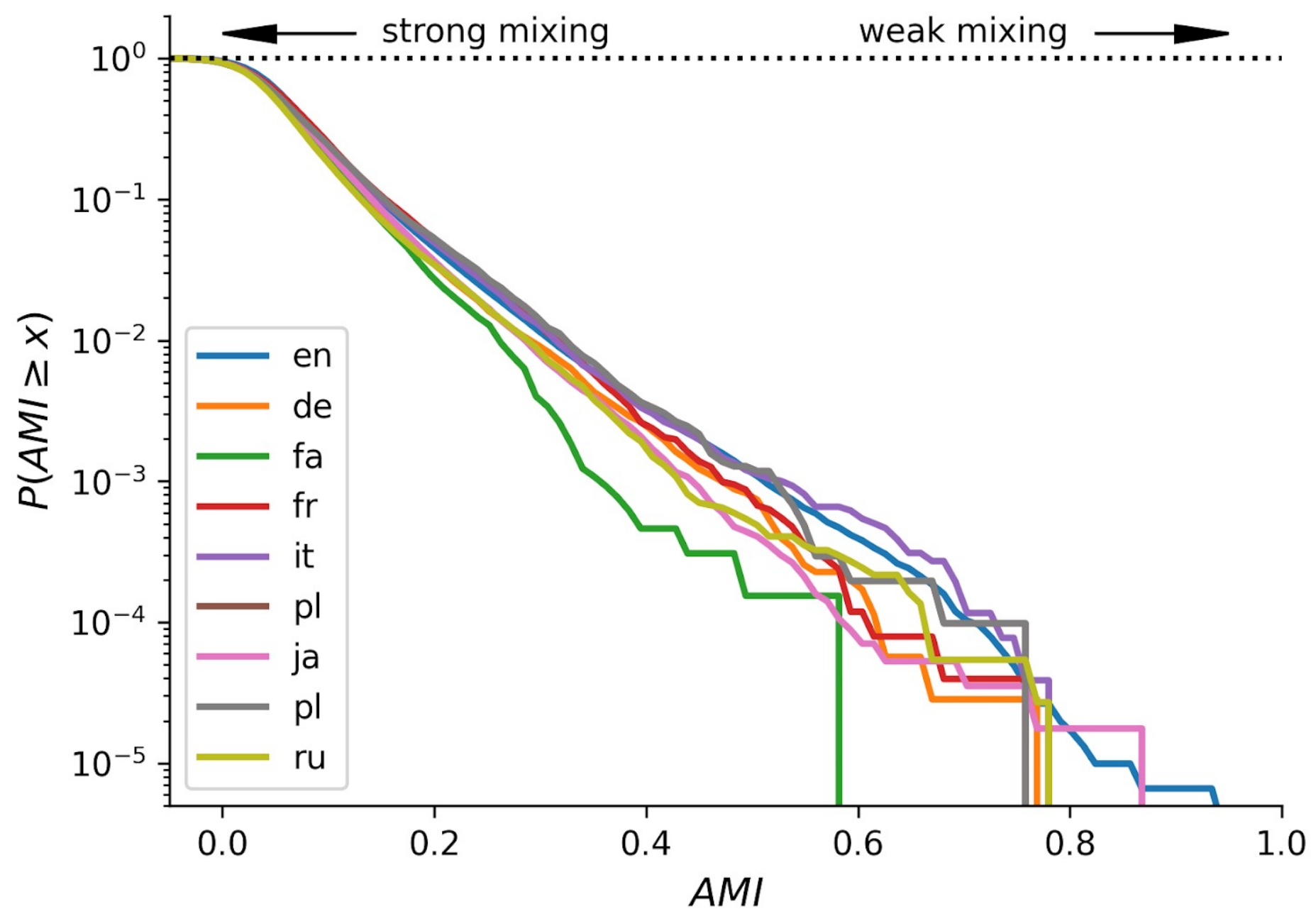
$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}.$$



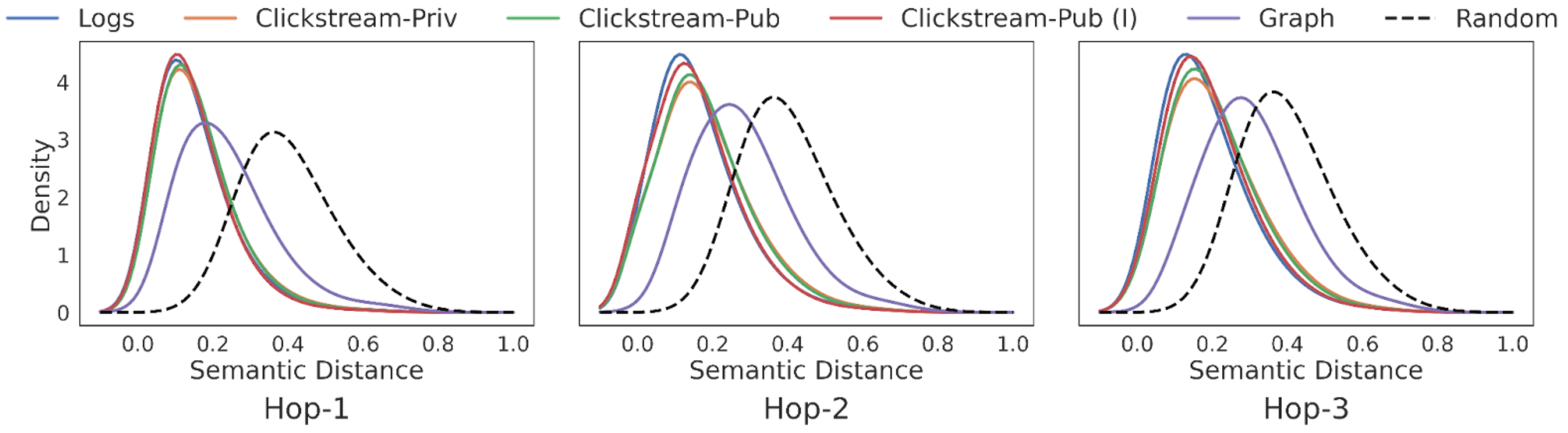
Strong mixing (AMI ≅ 0.1)

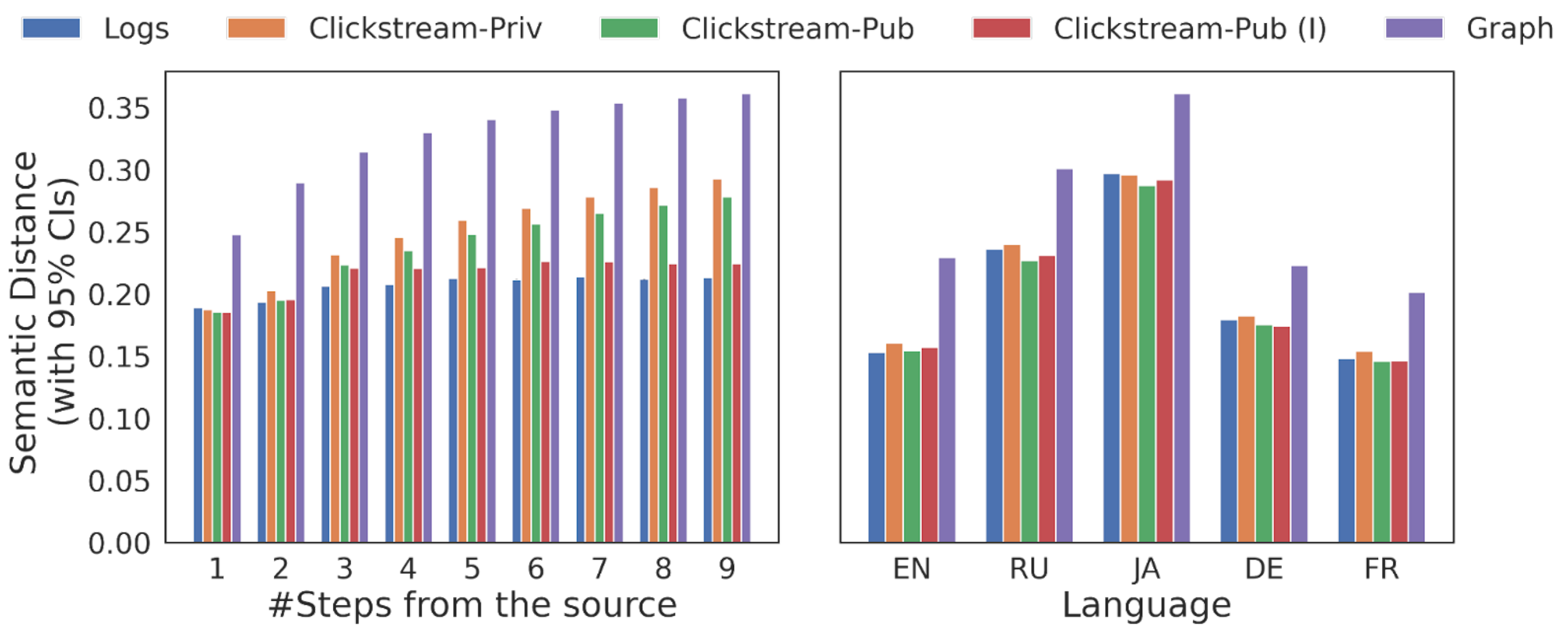
Weak mixing (AMI ≅ 0.6)

Majority of real trajectories exhibit strong mixing (AMI ≅ 0)



Diffusion in semantic space





Task	EN	JA	DE	RU	FR	IT	PL	FA
	Language version							
Semantic distance (k = 1)	-1.49	-0.98	-1.28	-1.25	-2.4	-2.33	-1.18	-0.79
Semantic distance (k = 3)	11.1	2.32	6.43	6.21	8.17	12.96	4.09	5.44
Semantic distance (k = 5)	28.93	5.12	19.11	14.77	19.3	36.43	9.24	14.91
Next-article prediction	9.20	8.85	8.32	8.60	8.86	9.93	7.58	3.64
Semantic relatedness	2.58	16.45	6.05	7.48	7.67	10.39	15.64	22.94
Semantic similarity	2.61	12.19	4.38	10.64	6.86	7.47	21.30	17.18
Topic classification	6.67	7.47	7.43	7.35	10.08	9.78	7.18	6.78
Link prediction (P@10)	-25.00	10.00	0.00	0.00	0.00	-11.11	-11.11	0.00
Link prediction (P@100)	-2.38	22.47	20.45	7.41	8.43	4.88	12.50	10.26

Differences are statistically significant but with 'small' (< 10%) effect sizes

Implications

For many cases, clickstream is good enough

- Research on navigation accessible to a wider audience
- User privacy: No need to store or reveal sensitive data!

Cases when clickstream is not good enough

- Tracking activities of the same user: revisitation patterns
- Reader interaction with additional content: e.g. images
- Understanding information consumption patterns

Broader Impact

Clickstream-like data can empower broader research on user navigation on online platforms

- Encouraging the community to release such datasets!

GitHub

PDF

W
Clickstream