Wikipedia Reader Navigation: When Synthetic Data is Enough*

Akhil Arora, Martin Gerlach, Tiziano Piccardi, Alberto García-Durán, Robert West

29th March

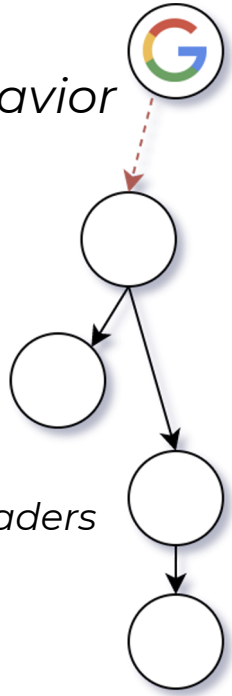
AMLD 2022

WIKIMEDIA FOUNDATION

* Meta page: https://w.wiki/4SQ4

Wikipedia reader navigation

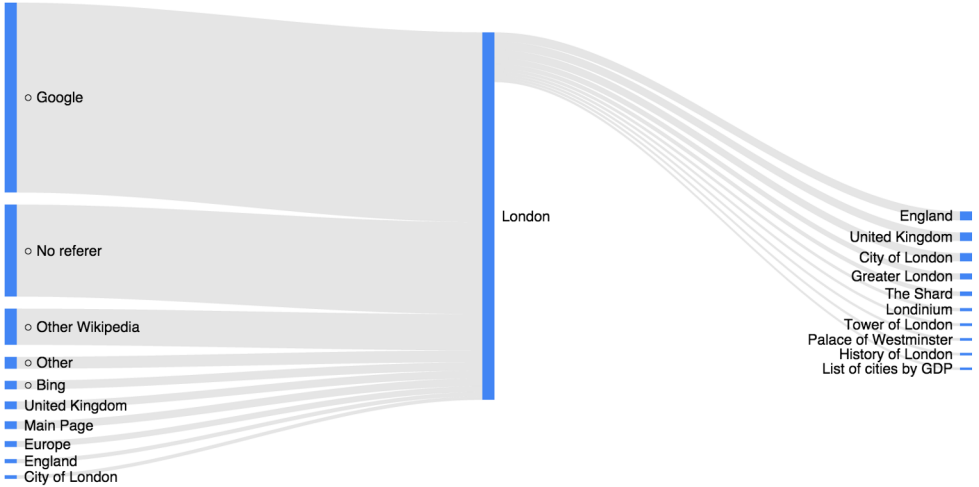
- *Readers leave “information-rich” traces of their navigation behavior*
- *Insights into navigation patterns have high utility*
  - *Understanding and better serving readers’ needs*
  - *Address knowledge gaps: identify missing/hard-to-find content*
  - *Identify and mitigate inherent structural biases (e.g. gender gaps)*
  - *Organize articles into a curriculum to improve the learning experience of readers*



- *Limited studies of reader navigation in Wikipedia*
- *Key challenge: Real navigation traces are kept private!*

Wikipedia Clickstream

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



- *Another challenge:*
  - *Next page visit depends only on the current page*
  - *Only captures first order navigation behavior*

Key research questions

- *How different are real trajectories from synthetic trajectories generated using the Wikipedia clickstream?*

- *How well can we approximate reader navigation via Wikipedia clickstream?*

Setup for evaluating ‘real*’ vs ‘synthetic’ trajectories

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.

Empirical characterization

- ***Mixing of flows***
- *Diffusion in semantic space*

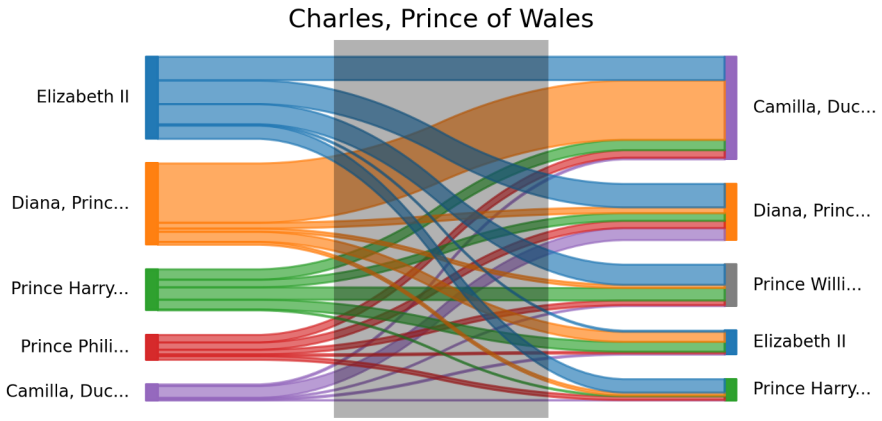
Downstream tasks

- ***Next-article prediction***
- *Link prediction*
- *Semantic relatedness*
- *Topic classification*

** Real trajectories are obtained from Webrequest server logs (include fingerprinting, and only ‘direct’ internal links)*

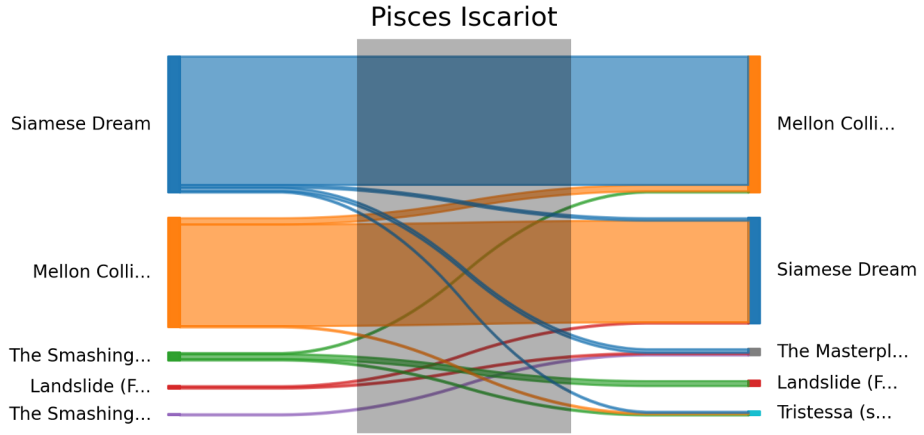
Mixing of Flows

Follow all trajectories passing through a given node. Connect source- and target-pages.



Strong mixing

(AMI ≅ 0.1)



Weak mixing

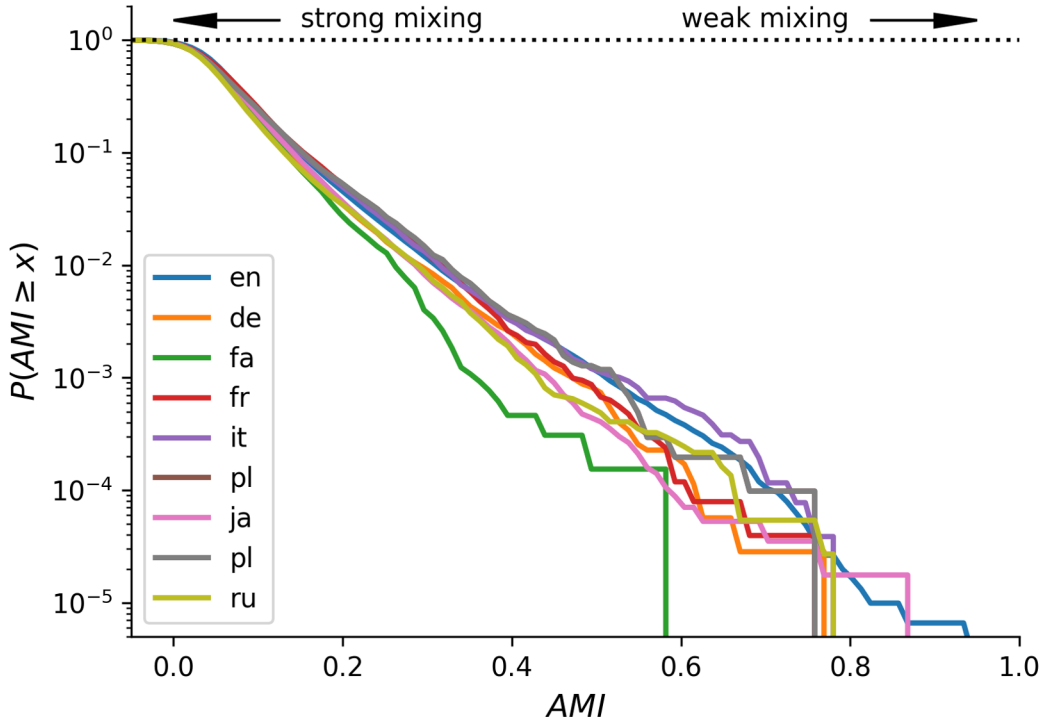
(AMI ≅ 0.6)

- *Quantify predictability using (adjusted) mutual information*

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

Stronger mixing (lower AMI) indicates low predictability

Mixing of Flows



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

Less than 10% pages have an AMI > 0.2

Next article prediction

Type	Dataset	All Queries								Filtered Queries							
		EN	JA	DE	RU	FR	IT	PL	FA	EN	JA	DE	RU	FR	IT	PL	FA
Real	LOGS	**0.369** †	**0.315** †	**0.275** †	**0.317** †	**0.316** †	**0.347** †	**0.302** †	**0.388** †	**0.595** †	**0.615** †	**0.646** †	**0.644** †	**0.690** †	**0.686** †	**0.693** †	**0.666** †
Synthetic	CLICKSTREAM-PRIV	0.325	0.273	0.249	0.286	0.279	0.307	0.277	0.361	0.541	0.557	0.593	0.587	0.625	0.618	0.634	0.623
Synthetic	CLICKSTREAM-PUB	0.316	0.258	0.238	0.259	0.266	0.278	0.247	0.270	0.541	0.561	0.592	0.589	0.629	0.618	0.641	0.642
Synthetic	CLICKSTREAM-PUB (I)	0.288	0.222	0.197	0.214	0.212	0.236	0.191	0.221	0.537	0.557	0.591	0.586	0.625	0.618	0.642	0.639
Synthetic	GRAPH	0.017	0.017	0.019	0.024	0.015	0.020	0.029	0.050	0.018	0.022	0.026	0.028	0.020	0.024	0.040	0.062

- *Train a markov order-2 model with input (s1, s2, t)*
- *Rank of true target (t* | s1,s2) in the ranked list obtained via P(t | s1,s2).*
- *Evaluate on a held-out test set using MRR*

- *Performance difference larger for low-resource languages*
  - *Hypothesis: k-anonymity (links > 10 clicks in Clickstream-Pub) plays a larger role than the restriction to first-order transitions*
  - *Difference mitigated in 'filtered' queries: prune all queries that lack observations in the training set*

†*Indicates statistical significance ($p < 0.05$) between the best and the second-best method using bootstrapped 95% confidence intervals*

Quantitative summary of relative differences (%) between Real and Synthetic Navigation sequences

Task	Language version							
	EN	JA	DE	RU	FR	IT	PL	FA
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

Takeaways

- *Real trajectories exhibit strong mixing (AMI ≅ 0)*
- *A small set of articles (≅ 0.1%) portrayed larger AMIs*
  - *Highlights cases where real trajectories differ substantially from synthetic*

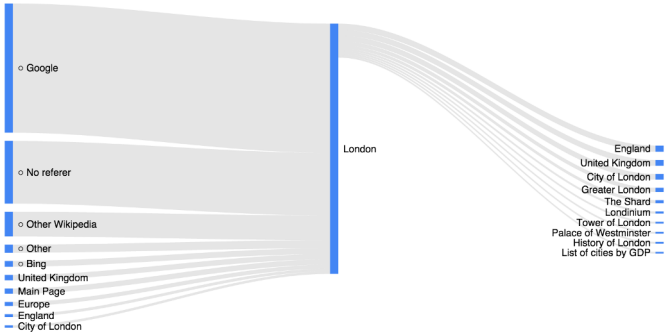
- *Clickstream data performance is within 10 % (or less) in comparison to real trajectories*
  - *Navigation embeddings from synthetic and real data are of comparable quality*

Quantitative evidence for the utility of Wikipedia clickstream as a public resource that can closely capture reader navigation on Wikipedia

Implications

- *For many cases, clickstream is good enough*
  - *Research on navigation in Wikipedia becomes accessible to a wider audience*
  - *User privacy: No need to store or reveal sensitive data!*

- *Cases exist, when real data is required (clickstream is not good enough)*
  - *Tracking activities of the same user: revisitation patterns, multi-tab behavior, etc.*
  - *How readers interact with additional content: images or infoboxes*
  - *Understanding information consumption patterns of Wikipedia readers*
- *Broader Impact*
  - *An open question whether our findings will generalize beyond Wikipedia*
  - *Clickstream-like data can empower broader research on user navigation on online platforms*
    - *Encouraging the community to release such datasets*



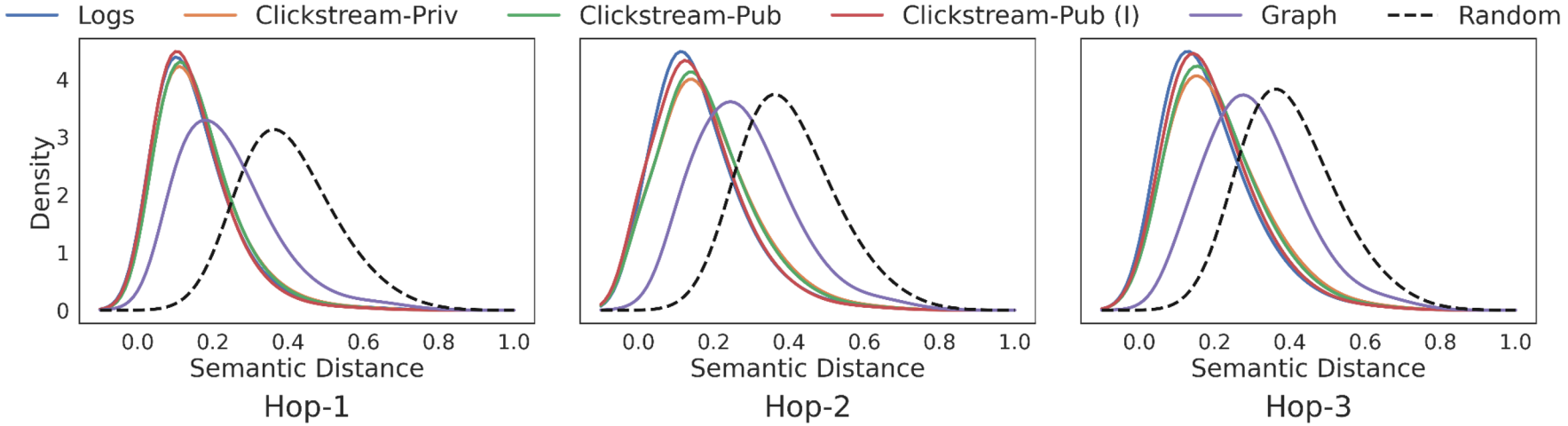
Thank you!

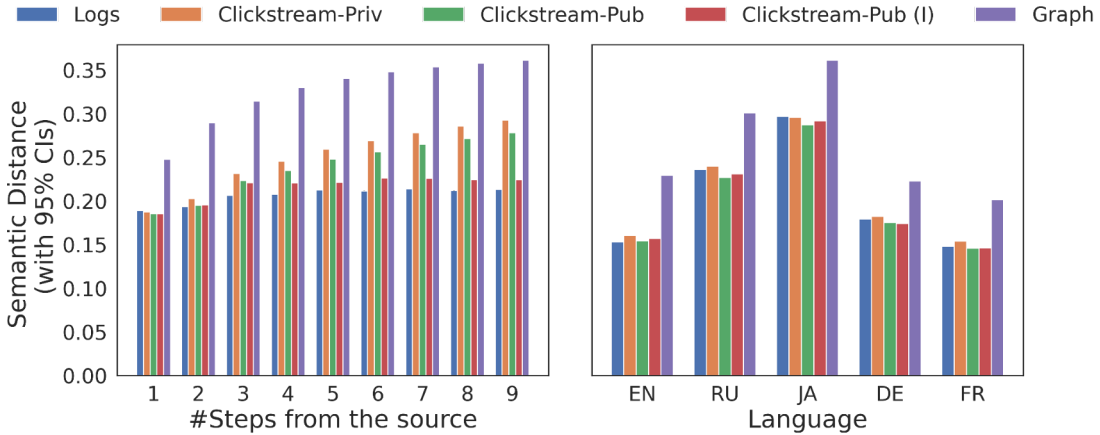
https://dlab.epfl.ch/people/aarora/

https://twitter.com/aroraakhilcs

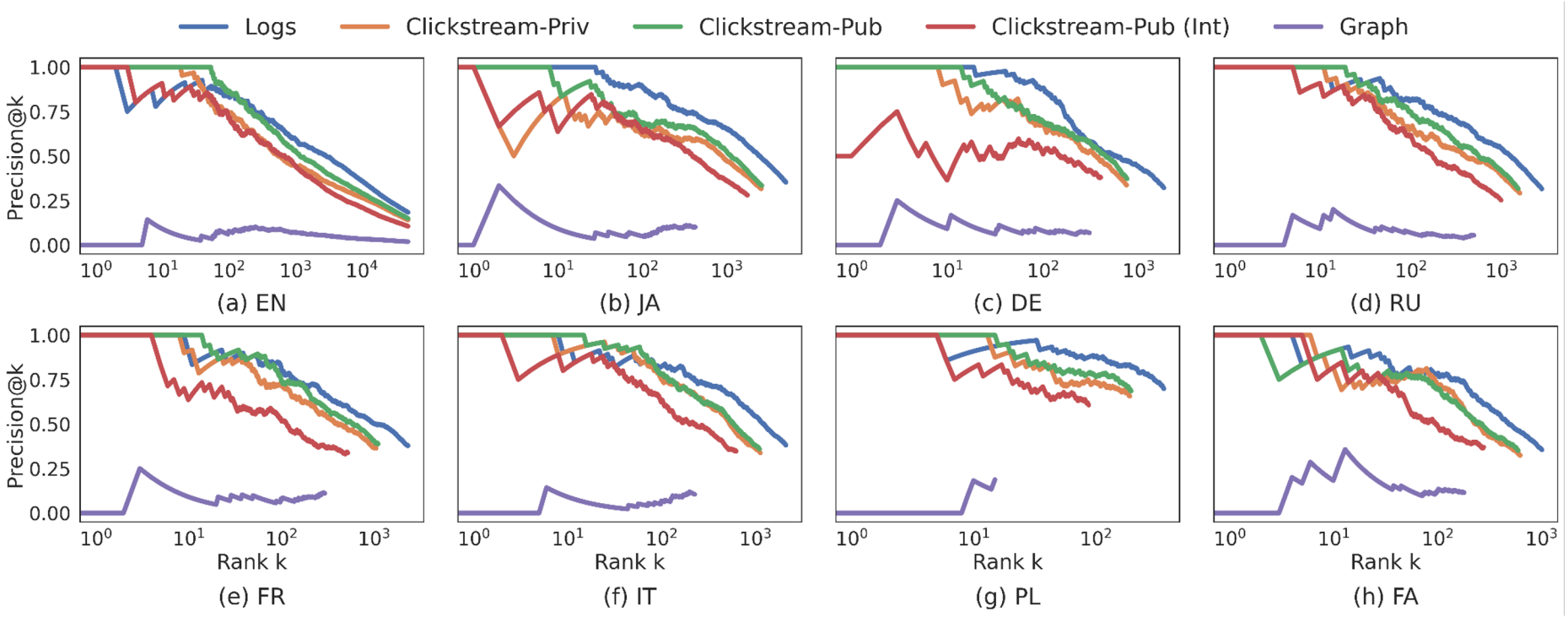
akhil.arora@epfl.ch

Diffusion in semantic space





Link prediction



Semantic relatedness & Topic classification

Spearman's rank correlation

		Relatedness								Similarity							
Type	**Dataset**	**EN**	**JA**	**DE**	**RU**	**FR**	**IT**	**PL**	**FA**	**EN**	**JA**	**DE**	**RU**	**FR**	**IT**	**PL**	**FA**
Real	Logs	0.769	0.728	0.693	**0.704**	0.697	**0.710**	0.691	0.665	0.722	**0.703**	0.648	**0.677**	0.662	0.665	0.630	0.621
Synthetic	Clickstream-Priv	0.764	0.689	0.673	0.688	0.714	0.703	**0.700**	0.595	0.711	0.662	0.633	0.623	**0.672**	0.652	**0.637**	0.539
Synthetic	Clickstream-Pub	0.749	0.625	0.653	0.655	0.647	0.643	0.597	0.541	0.703	0.626	0.621	0.612	0.620	0.619	0.520	0.530
Synthetic	Clickstream-Pub (I)	0.715	0.619	0.632	0.613	0.586	0.592	0.592	0.480	0.653	0.571	0.574	0.573	0.513	0.540	0.530	0.444
Synthetic	Graph	**0.771**	**0.750**	**0.709**	0.685	**0.723**	0.703	0.691	**0.677**	**0.734**	0.691	**0.674**	0.638	0.661	**0.666**	0.619	**0.633**

F1-score

		Micro Aggregates								Macro Aggregates							
Type	**Dataset**	**EN**	**JA**	**DE**	**RU**	**FR**	**IT**	**PL**	**FA**	**EN**	**JA**	**DE**	**RU**	**FR**	**IT**	**PL**	**FA**
Real	Logs	**0.628** †	**0.667**	0.621	**0.633** †	**0.618**	**0.623**	0.633	0.589	**0.569** †	**0.567** †	**0.547** †	**0.563** †	**0.560** †	**0.557** †	**0.541** †	0.496
Synthetic	Clickstream-Priv	0.597	0.646	0.595	0.609	0.589	0.594	0.609	0.571	0.544	0.547	0.523	0.539	0.532	0.531	0.512	0.477
Synthetic	Clickstream-Pub	0.586	0.618	0.575	0.586	0.556	0.562	0.587	0.549	0.531	0.513	0.491	0.506	0.496	0.489	0.478	0.446
Synthetic	Clickstream-Pub (I)	0.524	0.561	0.495	0.522	0.449	0.461	0.502	0.453	0.464	0.431	0.396	0.436	0.378	0.375	0.387	0.335
Synthetic	Graph	0.625	0.666	**0.636** †	0.628	**0.625** †	0.621	**0.639**	**0.600** †	0.563	0.543	0.535	0.547	0.555	0.543	0.526	**0.499**

†*Indicates statistical significance ($p < 0.05$) between the best and the second-best method using bootstrapped 95% confidence intervals*