node2bits: Compact Time- and Attribute-aware Node Representations for User Stitching

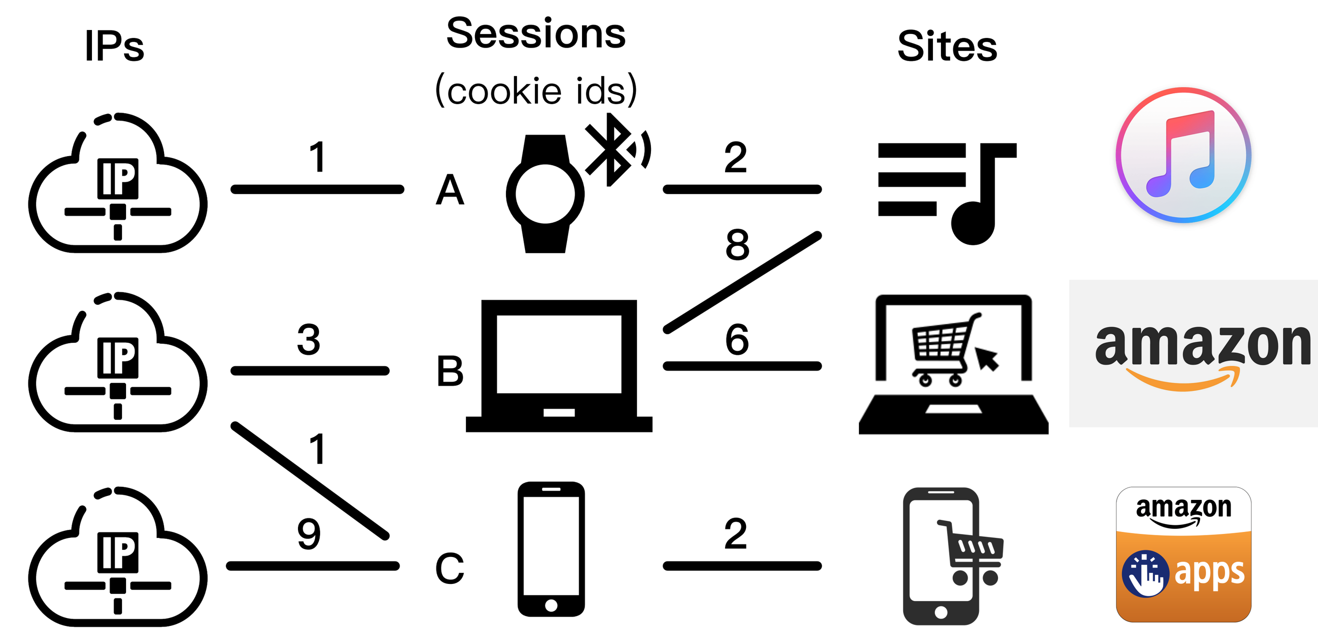
Di Jin, Mark Heimann , Ryan A. Rossi, Danai Koutra

Problem: Identity stitching

- Identifying and matching various online references to the same user in real-world web services.
- Applies to large-scale web data with limited access to textual info
- Crucial to personalization & recommendation

Challenges. Network embedding approach to identity stitching:

1) Graph heterogeneity
2) Temporal dynamics
   i. Temporal validity
   ii. Functional similarity
3) Quadratic comparison in similarity search
4) Storage inefficiency



Web device graph data

We propose node2bits, an efficient framework that captures *temporal dynamics* from a *heterogeneous* interaction network into *sparse binary embeddings* to perform identity stitching.

① Temporal random walk & context

- Temporal random walk: a sequence of nodes connected by edges with non-decreasing timestamps.
  - Invalid walk: f → c → d ✗
  - Valid walk: a → b → c → d ✓
    - Nodes along the walk are temporally valid.
- Temporal locality: controlled by random walk strategies.
  - *Uniform*: unbiased temporal random walk (RW)
  - *Short term*: bias towards edges close in time.
  - *Long term*: bias towards edges late in time (e.g., i → j)

$$P(v|u) = \frac{e^{\tau_{u,v}}}{\sum_{w \in \Gamma_u} e^{\tau_{u,w}}}$$ transition probability

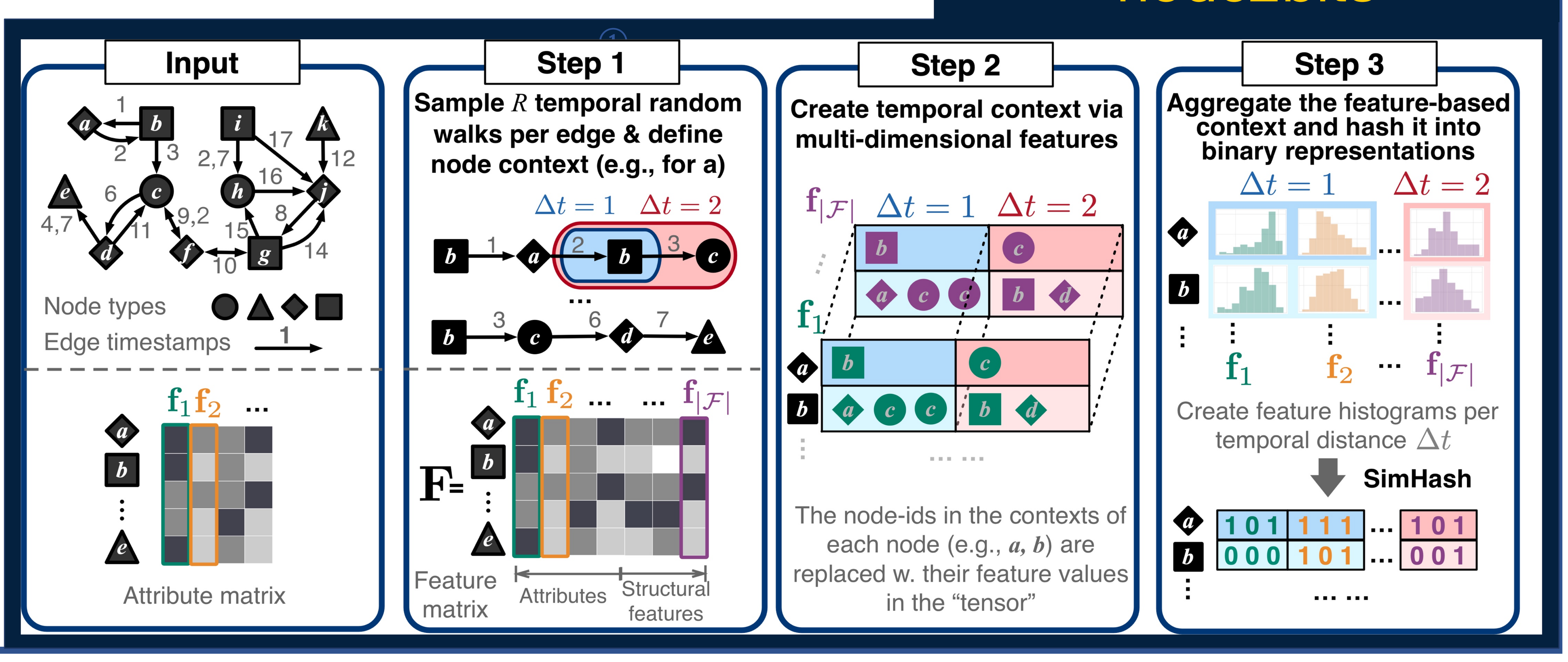
② Multi-dimensional feature temporal context

- Functional similarity: measured via the histograms of typed entities across temporal distances.
- e.g., 1-hop context of *a* is {*b*}, 2-hop context is {*c*}
  - Augment contexts from multiple walks.
- Histograms
  - Sparsity for computational efficiency
  - Less lossy: both strong & weak values preserved

③ Feature aggregation & Hashing

- Aggregation
  - Iterate over all feature distributions across typed contexts
- SimHash the histograms
  - Derive sparse binary hashcodes using sketching

node2bits



Evaluation

① Identity stitching (supervised) on static graphs

	Metric	CN	SE	LINE	DW	n2vec	s2vec	DNGR	m2vec	AspEm	N2B-0
citeseer	**AUC**	0.9141	0.4846	0.5481	0.5614	0.6188	0.9344	0.5015	0.5546	0.5049	0.9480*
	ACC	0.9141	0.5045	0.5372	0.5579	0.6211	0.8936	0.4688	0.5357	0.5223	0.9196*
	F1	0.9137	0.5028	0.5371	0.5547	0.6159	0.8926	0.4682	0.5348	0.5222	0.9192*
yahoo	**AUC**	0.6851	0.5378	0.8050	0.7640	0.7636	OOT	OOM	0.8233	0.4938	0.8088
	ACC	0.6851	0.4760	0.7771	0.7117	0.7233			0.7827	0.5018	0.8010
	F1	0.6505	0.4375	0.7764	0.7117	0.7231			0.7823	0.5018	0.7987

Identity stitching (supervised) on temporal graphs

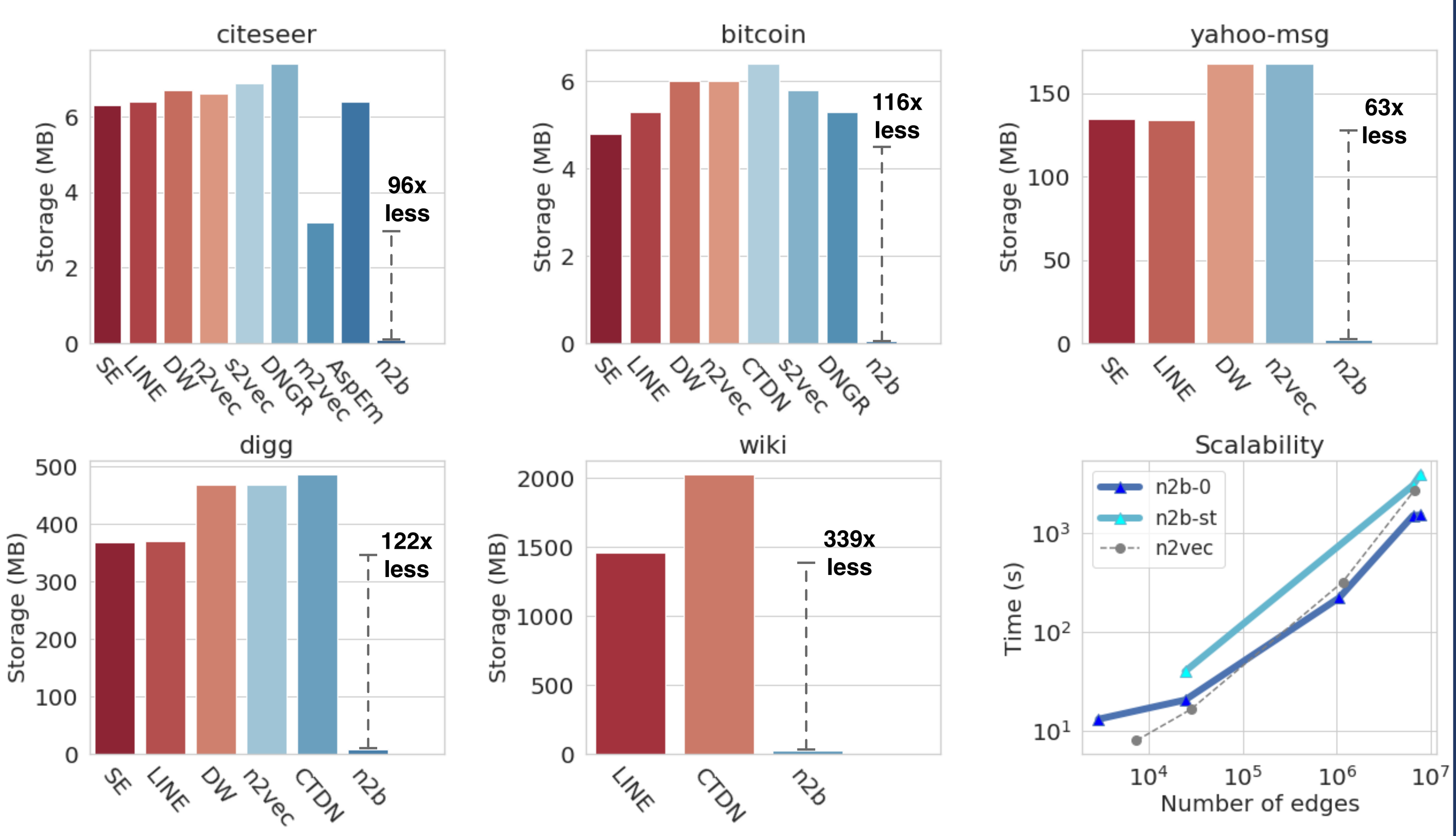
	Metric	CN	SE	LINE	DW	n2vec	s2vec	DNGR	AspEm	CTDNE	N2B-0	N2B-SH	N2B-LN
bitcoin	**AUC**	0.7474	0.5828	0.6071	0.6306	0.6462	0.8025	0.5909	0.5344	0.6987	0.7584	0.7609	0.7380
	ACC	0.7174	0.5842	0.5842	0.6158	0.6158	0.7263	0.5526	0.5316	0.6000	0.7211	0.7268	0.6737
	F1	0.7001	0.5728	0.5828	0.6158	0.6157	0.7263	0.5525	0.5315	0.5964	0.7209	0.7271	0.6735
digg	**AUC**	0.6217	0.5171	0.7878	0.7398	0.7445	OOT	OOM	0.5105	0.6967	0.8185*	0.7611	0.7587
	ACC	0.6217	0.5152	0.7694	0.6971	0.7013			0.5088	0.5915	0.7982*	0.7418	0.7444
	F1	0.5585	0.3770	0.7683	0.6960	0.7003			0.5088	0.5884	0.7958*	0.7411	0.7433
wiki	**AUC**	0.6997	OOT	0.7854	OOM	OOM	OOT	OOM	0.5374	0.7707	0.8230	0.8259*	0.8214
	ACC	0.6997		0.7132					0.5141	0.6488	0.7145	0.7510*	0.7103
	F1	0.6699		0.7129					0.5141	0.6398	0.7088	0.7476*	0.7067
comp-X	**AUC**	0.5970	OOM	0.5000	OOM	OOM	OOT	OOM	0.5213	OOM	0.8095*	0.7496	0.7525
	ACC	0.5970		0.6757					0.5103		0.8414*	0.7959	0.7975
	F1	0.5189		0.4032					0.5103		0.8154*	0.7581	0.7606

Identity stitching (unsupervised) on static & temporal graphs

Metric	citeseer		yahoo		bitcoin		digg		wiki	
	CN	N2B-U	CN	N2B-U	CN	N2B-U	CN	N2B-U	CN	N2B-U
ACC	0.9141	0.8661	0.6851	0.7553	0.7474	0.7684	0.6217	0.7157	0.6997	0.7350
F1	0.9137	0.8660	0.6505	0.7518	0.7301	0.7663	0.5585	0.7074	0.6699	0.7349

② Output storage. 63 – 339 x less than baselines.

③ Scalability. node2bits scales well with the graph size.



References

- Bhattacharya, I., Getoor, L.: Collective entity resolution in relational data. TKDD 1 (1), 1–36 (2007)
- Dasgupta, A., Gurevich, M., Zhang, L., Tseng, B., Thomas, A.O.: Overcoming browser cookie churn with clustering. In: WSDM. pp. 83–92 (2012)
- Nguyen, G H, Boaz Lee, J..: Dynamic Network Embeddings: From Random Walks to Temporal Random Walks. In: BigData. (2018)