Vidisha Bhat & Devin Tark

Growing Attributed Networks through Local Processes

Harshay Shah
University of Illinois at
Urbana-Champaign, IL, USA
hrshah4@illinois.edu

Suhansanu Kumar
University of Illinois at
Urbana-Champaign, IL, USA
skumar56@illinois.edu

Hari Sundaram
University of Illinois at
Urbana-Champaign, IL, USA
hs1@illinois.edu

Question: How do we form networks?

Previous models rely on:

1. Preferential Attachment & Fitness

An author finding citations for a paper would pick papers from the entire network in proportion to their in-degree or fitness.

2. Vertex Copying

Or, would pick a paper uniformly at random from *all* papers and either cite or copy its citations.

What's the problem with this?

In the real world, individuals form networks with limited information and partial network access

Previous Models - Related Work

- Preferential Attachment and Fitness
  - New node links to an existing node with probability proportional to degree or fitness of the new node.
  - Only consider power law degree distribution and small diameter.
  - Additional mechanisms are needed for other properties like clustering and attribute mixing patterns.
  - DMS, RL and KA models.
- Triangle Closing
  - Form edges with nodes having more common neighbors.
  - Increases local clustering.
  - Do not preserve the local clustering of low degree nodes.

Papers :

A preferential attachment paradox: How preferential attachment combines with Growth to produce networks with Log Normal in degree distribution

Previous Models - Related Work Cont.

- Attributed Network Models
  - Models account for the effect of attribute homophily on edge formation and preserve mixing patterns.
  - Fitness based models and Microscopic models.
  - However models like SAN and KA preserve assortative mixing patterns and degree distribution but not local clustering and degree-clustering correlation.
- Random Walk Models
  - Inherently local.
  - Generate networks with power law degree distribution and small diameter, but not clustering.
  - SK and HZ incorporate triadic closure but not skewed local clustering of real world networks.
  - Models like FF overestimate local clustering and degree clustering relationship of real world networks.
  - Existing models also disregard the effect of homophily and attribute mixing patterns.

The existing models don’t explain how resource constraints and local processes jointly preserve multiple global properties of attributed networks.

Goal: Develop a growth model that accounts for:

A. Resource constraints
B. Unifying sociological phenomena that inform how we develop networks:
    a. Bounded Rationality
    b. Structural Constraints
    c. Triadic Closure
    d. Attribute Homophily
    e. Preferential Attachment

Proposal: The Attributed Random Walk (ARW) model. Incoming nodes select a seed node based on attribute similarity, and initiate a ***biased*** random walk

- **Bounded Rationality:** Individuals are boundedly rational and have resource constraints, we employ simple rules to determine forming edges in the face of limited info and partial network access.
- **Structural Constraints:** Network distance acts as a constraint that limits long range connection
- **Triadic Closure:** Nodes with common neighbors have increased likelihood of forming edges, e.g. your friends becoming friends
- **Attribute Homophily:** Nodes that have similar attributes are more likely to form links
- **Preferential Attachment:** Nodes tend to link to high degree nodes that have a lot of visibility., e.g. if you see a paper being cited a lot you’re probably going to cite that paper.

Problem: Given a reference network $G = (V,E,B)$ where V and E are the set of nodes and edges respectively, and nodes have attributes $b \in B$

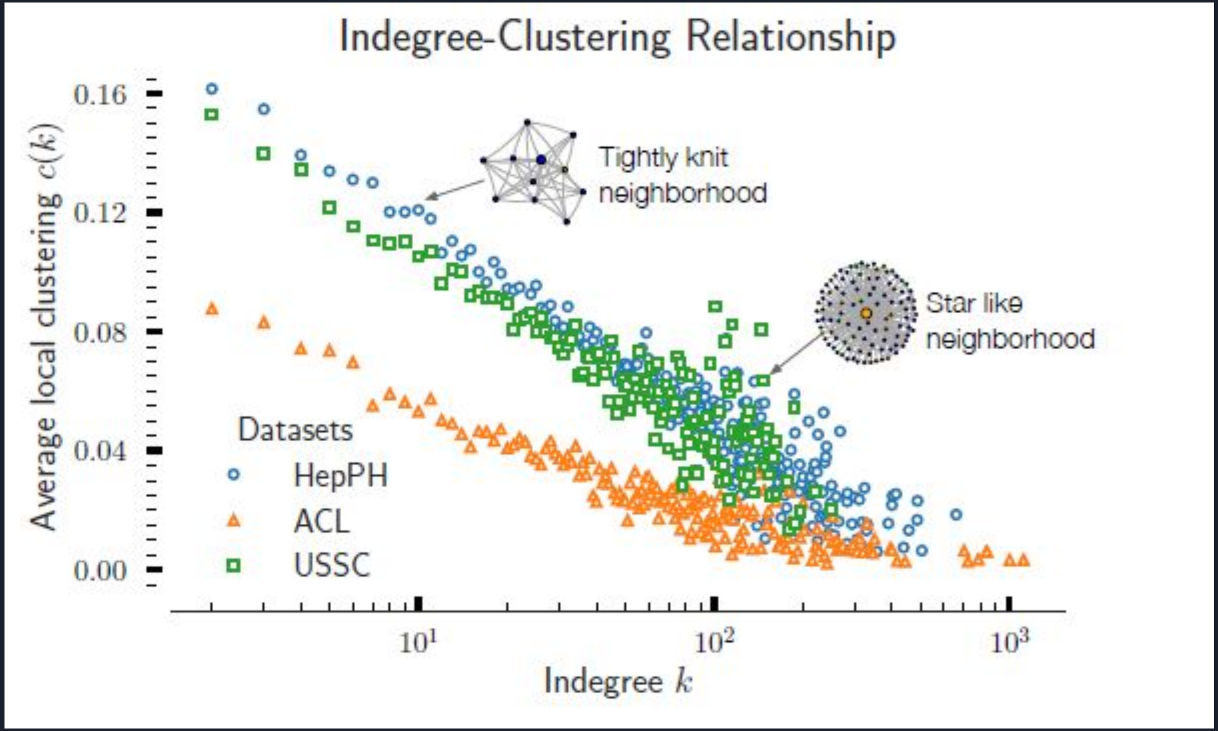
Growth model should be:

- ★ Normal
  - Account for sociological phenomena that influences how individuals form edges under limited global information and partial network access
- ★ Accurate
  - Preserve key structural and attribute based properties: degree distribution, local clustering, degree-clustering relationship and attribute mixing patterns.
- ★ Parsimonious
  - Tunable by a few parameters

Global Properties of Networks

- Degree Distribution
  - Real-world networks tend to exhibit heavy-tailed degree distribution where a small but significant number of nodes are high degree hubs
- Local Clustering
  - Real-world network exhibit high local clustering. Influenced by triadic closure where nodes that share neighbors are more likely to form a connection
- Homophily
  - Nodes with similar attributes are more likely to form links with each other than with those with different attributes.
- Increasing out-degree over time
  - The out-degree of nodes that join networks tends to increase over time. This densifies networks and follows a power law relationship where e(t) is proportional to n(t) to some constant power.

Can these global properties be preserved with only local processes?



Indegree-Clustering Relationship
Average local clustering $c(k)$
0.16
0.12
0.08
0.04
0.00
Tightly knit neighborhood
Star like neighborhood
Datasets
HepPH
ACL
USSC
10^1
10^2
10^3
Indegree k

Enter the Attributed Random Walk Model

RESEARCH-ARTICLE

News labeling as early as possible: real or fake?

Authors: Maryam Ramezani, Mina Rafiei, Soroush Omranpour, Hamid R. Rabiee

Authors Info & Affiliations

Publication: ASONAM '19: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining • August 2019 • Pages 536–537 • https://doi-org.proxy2.library.illinois.edu/10.1145/3341161.3342957

ABSTRACT

Differentiating between real and fake news propagation through online social networks is an important issue in many applications. The time gap between the news release time and detection of its label is a significant step towards broadcasting the real information and avoiding the fake. Therefore, one of the challenging tasks in this area is to identify fake and real news in early stages of propagation. However, there is a tradeoff between minimizing the time gap and maximizing accuracy. Despite recent efforts in detection of fake news, there has been no significant work that explicitly incorporates early detection in its model. The

real news in early stages of propagation. However, there is a tradeoff between minimizing the time gap and maximizing accuracy. Despite recent efforts in detection of fake news, there has been no significant work that explicitly incorporates early detection in its model. The

proposed method utilizes recurrent neural networks with a novel loss function, and a new stopping rule. Experiments on real datasets demonstrate the effectiveness of our model both in terms of early labelling and accuracy, compared to the state of the art baseline and models.

References

1. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *SIGKDD Explor. Newsl.*, vol. 19, no. 1, sep 2017. Open URL

2. Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, pp. II-1188--II-1196. Open URL

3. Y. Liu and Y.-f. B. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *Proceedings of the Conference on Artificial Intelligence*, 2018, pp. 354--361. Open URL

4. J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2016, pp. 3818--3824. Open URL

5. N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," in *Proceedings of the Conference on Information and Knowledge Management*, 2017, pp. 797--806. Open URL

. arXiv preprint

reprint

ediction with guided

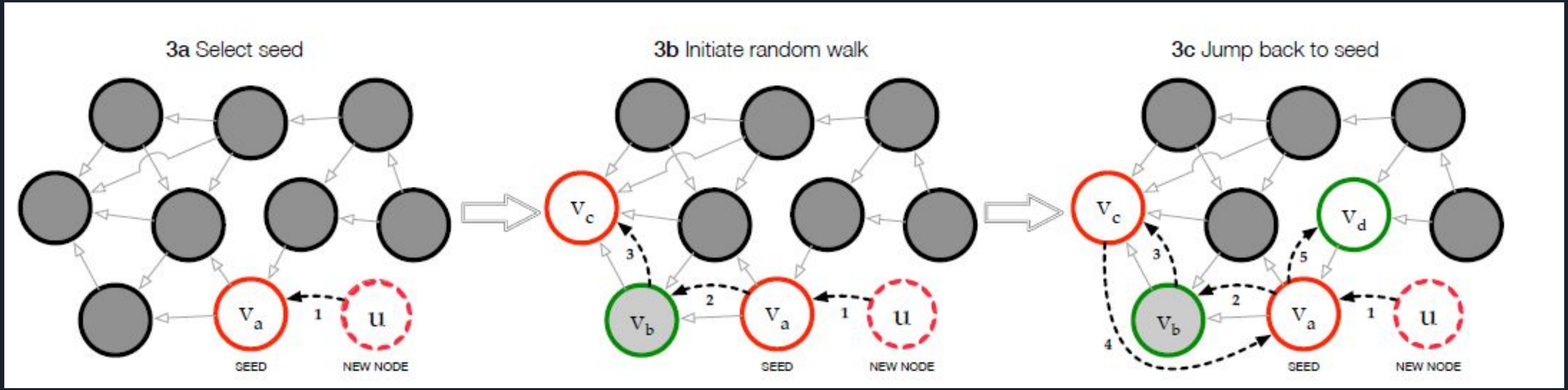
L .

ontent-based

Detection on

cial Attention

cking: Detecting



Select-Seed

(1) With probability $p_{\text{same}}/p_{\text{same}}+p_{\text{diff}}$, randomly select a seed node from existing nodes that have the same attribute value, $B(u)$.

(2) Otherwise, with probability $p_{\text{diff}}/p_{\text{same}}+p_{\text{diff}}$, randomly select a seed node from existing nodes that do *not* have the same attribute value, $B(u)$.

Random-Walk

(1) At each step of the walk, new node u visits node v_i.

- If $B(u) = B(v_i)$, u links to v_i with probability p_{same}
- Otherwise, u links to v_i with probability p_{diff}

(2) Then, with probability p_{jump}, u jumps back to seed s_u.

(3) Otherwise, with probability $1 - p_{\text{jump}}$, u continues to walk. It picks an outgoing edge with prob. p_{out} *or* an incoming edge with prob. $1 - p_{\text{out}}$ to visit a neighbor of v_i.

(4) Steps 1-3 are repeated until u links to $m(t)$ nodes.

Parameterizing Normal Behavior in ARW

- Phenomenon 1: Limited Resources
  - ARW only requires node level information about the immediate 1-hop neighborhood of a node
- Phenomenon 2: Structural Constraints
  - A Random Walk based model, p_{jump} controls the probability that the walker jumps back to the seed node, limiting distance
- Phenomenon 3: Triadic Closure
  - In attribute-less network, a node will link to another with probability p_{link} and triad with be closed proportional to p^2_{link}
- Phenomenon 4: Attribute Homophily
  - Can tune with p_{same} and p_{diff}
- Phenomenon 5: Preferential Attachment
  - Controlled with P_{out} - probability of traversing an outgoing vs incoming edge. Random walks tend to visit nodes with high indegree, but can be tuned down by lowering P_{out} towards 0.

Experiments

Compare G and G' on four properties:

1. Degree Distribution
2. Local clustering distribution
3. Degree clustering relationship
4. Attribute Assortativity

Evaluation Metric :

Weighted Relative Error (WRE) for degree-clustering relationship.

Model	Abbreviation	Type	Attributed?
Dorogovtsev et al. [13]	DMS	PA	✗
Relay Linking [41]	RL	PA	✗
Kim-Altmann [22]	KA	PA	✓
Social Attribute Network [17]	SAN	PA+TC	✓
Holme-Kim [20]	HK	PA+TC	✗
Herera-Zufiria [19]	HZ	RW	✗
Saramaki-Kaski [37]	SK	RW	✗
Forest Fire [27]	FF	RW	✗

Results

Significance level: p < 0.001 (dark), p < 0.01 (light)

		A: Indegree Distribution (KS Stat) USSC	HepPH	Semantic	ACL	APS	Patents	B: Local Clustering Distribution (KS Stat) USSC	HepPH	Semantic	ACL	APS	Patents	C: Indegree & Clustering Relationship (WRE) USSC	HepPH	Semantic	ACL	APS	Patents	Assortativity $\lvert r-\hat{r}\rvert < \epsilon$
Preferential Attachment	DMS	0.03	0.03	0.05	0.09	0.04	0.02	0.80	0.82	0.56	0.63	0.83	0.50	1.00	1.00	1.00	1.00	1.00	1.00	✗
Preferential Attachment	KA	0.11	0.19	0.22	0.26	0.13	0.06	0.80	0.82	0.56	0.63	0.82	0.50	1.00	1.00	1.00	1.00	1.00	1.00	✓
Preferential Attachment	RL	0.12	0.12	0.17	0.15	0.07	0.15	0.79	0.82	0.56	0.62	0.83	0.50	0.99	1.00	1.00	0.99	1.00	1.00	✗
Triangle Closing	HK	0.11	0.19	0.22	0.26	0.13	0.05	0.39	0.55	0.15	0.08	0.52	0.05	0.59	0.74	0.08	0.25	0.73	0.17	✗
Triangle Closing	SAN	0.12	0.18	0.19	0.24	0.11	0.05	0.12	0.05	0.12	0.16	0.05	0.19	0.13	0.14	0.34	0.31	0.15	1.28	✓
Random Walk	FF	0.16	0.17	0.14	0.12	0.46	0.32	0.53	0.54	0.33	0.69	0.19	0.40	1.64	1.74	0.54	4.11	0.15	0.73	✗
Random Walk	SK	0.19	0.22	0.25	0.27	0.13	0.13	0.15	0.29	0.26	0.34	0.34	0.11	0.14	0.46	0.74	0.41	0.51	0.38	✗
Random Walk	HZ	0.18	0.22	0.23	0.26	0.13	0.13	0.08	0.29	0.10	0.07	0.34	0.03	0.18	0.45	0.21	0.22	0.51	0.04	✗
Random Walk	**ARW**	0.07	0.06	0.07	0.09	0.07	0.08	0.08	0.04	0.05	0.05	0.05	0.09	0.14	0.10	0.05	0.13	0.08	0.08	✓

Results Highlights

- ★ Preferential attachment models (DMS RL KA) preserve in-degree distribution, but not clustering because they don't account for triadic closure

- ★ Models that use triangle closing (HK SAN) lead to considerable improvement in local clustering but perform poorly with degree-clustering relationship

- ★ Existing random walk models FF SK HZ do not account for homophily and attribute mixing patterns.

Some Limitations

Network	Description	$\|V\|$	$\|E\|$	T	A, $\|A\|$	LN (μ, σ)	DPL α	Avg. LCC	AA r
USSC [14]	U.S. Supreme Court cases	30,288	216,738	1754-2002	-	(1.19, 1.18)	2.32	0.12	-
HEP-PH [15]	ArXiv Physics manuscripts	34,546	421,533	1992-2002	-	(1.32, 1.41)	1.67	0.12	-
Semantic [2]	Academic Search Engine	7,706,506	59,079,055	1991-2016	-	(1.78, 0.96)	1.58	0.06	-
ACL [36]	NLP papers	18,665	115,311	1965-2016	VENUE, 50	(1.93, 1.38)	1.43	0.07	0.07
APS [1]	Physics journals	577,046	6,967,873	1893-2015	JOURNAL, 13	(1.62, 1.20)	1.26	0.11	0.44
Patents [27]	U.S. NBER patents	3,923,922	16,522,438	1975-1999	CATEGORY, 6	(1.10, 1.01)	1.94	0.04	0.72

- ★ Data set?
- ★ Attribute Similarity?

Harshay Shah, Suhansanu Kumar, Hari Sundaram. 2017. Growing Attributed Networks through Local Processes. arXiv preprint arXiv:1712.10195 (2017)