# Second Space: A Generative Model for the Blogosphere

**Amit Karandikar, Akshay Java, Anupam Joshi, Tim Finin, Yaacov Yesha and Yelena Yesha**
University of Maryland, Baltimore County, Baltimore MD 21250
{amitk1,aks1,joshi,finin,yayesha,yeyesha}@cs.umbc.edu

Analysing complex natural phenomena often requires synthesized data that matches observed characteristics. Graph models are widely used in analyzing the Web in general, but are less suitable for modeling the Blogosphere. While blog networks resemble many properties of Web graphs, the dynamic nature of the Blogosphere, its unique structure and the evolution of the link structure due to blog readership and social interactions are not captured by the existing models. We describe an agent-based simulation model to construct blog graphs that exhibit properties similar to the real world blog networks in their degree distributions, degree correlation, clustering coefficient and reciprocity. The model can help researchers analyze the Blogosphere and facilitates the development and testing of new algorithms.

We capture the linking patterns arising in the Blogosphere through *local interactions*, as suggested in (Vazquez 2003). These local interactions comprise the interactions of a blog with other blogs connected to it either by inlinks or outlinks. Key properties of the resulting graph include the degree distributions, degree correlation, clustering coefficient, average degree, reciprocity and the distribution of connected components. To the best of our knowledge, there exist no general models to generate the blog and post networks which possess the properties observed in the real world blogs. Table 1 gives a quick comparison of the existing Web model and shows the need for a model for Blogosphere.

**Design.** The observed characteristics of bloggers effect the structure of the two graphs underlying the Blogosphere as shown in Figure 1. The *blog network* is a network of blogs obtained by collapsing all directed post links between blog posts into directed edges between blogs. Blog networks give a macroscopic view of the Blogosphere and help to infer a social network structure, under the assumption that blogs that are part of the same community link each other more often. The *post network* (Leskovec *et al.* 2007) is formed by ignoring the posts' parent blogs and focusing on the link structure among posts only. Post networks give a microscopic view of the Blogosphere with details like which post linked to which other post and at what time.

The Pew Internet and American Life Project (Lenhart & Fox 2006) survey found that (i) blog writers are enthusiastic blog readers, (ii) most bloggers post infrequently, and (iii) blog readership can be inferred using blogrolls. Active bloggers are more likely to have a blogroll and follow it regularly. We built these observations into our model through the following aspects using six features described in (Karandikar *et al.* 2007).

Copyright © 2008, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

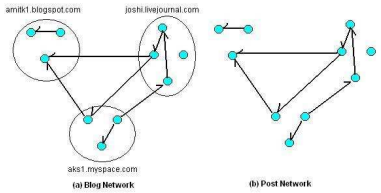


Figure 1: The Blogosphere comprises two graphs: a blog network and post network.

| Symbol | Explanation |
|---|---|
| **rR** | Probability of *random reads* |
| **rW** | Probability that *writers* are selected at *random* |
| **pD** | Probability that new nodes *don't connect* to network |
| **g** | Growth function exponent |
| $ts_i$ | $i^{th}$ step of the graph evolution |
| $RM_j$ | Blogger *j*'s read memory (finite FIFO queue) |
| M | Initial number of blog nodes |
| N | Total blog nodes to create |
| *p(k, j)* | $k^{th}$ post of a blog node $b_j$ |
| *e(t)* | Expected number of edges at step *t* (by growth function) |

Table 2: A small number of key parameters control the second space model.

The model proposed in (Pennock *et al.* 2002) captures the random behavior but does not capture the local interactions among nodes in the graph. We use the "alpha preferential attachment" model proposed in (Chung & Lu 2006) to obtain power law degree distributions in a directed graph. We have modified this model to reflect local interactions among the bloggers by *changing preferential attachment among all existing nodes to just neighbors of a node*. The main input parameters for the model are given in **bold** in Table 2.

We define our model using four main parameters, $rR$, $rW$, $pD$ and $g$, which are adjusted to vary the properties of the generated graphs. In the network initialization step, we start with M $(M << N)$ blog nodes and blog-post graph

| Property | ER model | BA model | Blogosphere | Simulation |
|---|---|---|---|---|
| Type | undirected | undirected | **directed** | directed |
| Degree distribution | poisson | power law | power law | power law |
| Slope [inlinks,outlinks] | N/A | [2.08,-] | [1.66-1.8,1.6-1.75] | [1.7-2.1,1.5-1.6] |
| Avg. degree | constant (for given p) | constant (adds m edges) | **increases** | increases |
| Component distribution | N/A (undirected) | N/A (undirected) | **Power law** | Power law |
| Correlation coefficient | - | 1 (fully preferential) | **0.024** (WWE) | 0.1 |
| Avg clustering coeff. | 0.00017 | 0.00018 | **0.0235** (WWE) | 0.0242 |
| Reciprocity | N/A (undirected) | N/A (undirected) | **0.6** (WWE) | 0.6 |

Table 1: Comparing the properties of common Web models, the observed Blogosphere and our simulated Blogosphere reveals key differences.

G(V, E) such that $V = M$, $E = \{\}$. The read queue of all blog nodes is empty at the start and fills as the bloggers read existing blog-posts in the read phase. In each cycle of the simulation, we perform steps in which simulated bloggers read blogs until motivated to write and then write commenting on one of the blogs recently read.

**Results.** We compared the properties of our simulated Blogosphere with the properties of two large blog datasets available for researchers – WWE 2006 and ICWSM 2007. Tables 3 and 4 summarize the properties of ICWSM and WWE datasets and compare them with the simulated Blogosphere. We eliminated spam blogs from WWE dataset using the techniques from (Kolari *et al.* 2006). The WWE dataset was largely biased toward LiveJournal, MySpace and few other blogs. Hence we ignored all post links to and from these blogs. A complete interpretation and analysis of the results is available in (Karandikar *et al.* 2007).

| Blog network | ICWSM | WWE | Simulation |
|---|---|---|---|
| Total blogs | 159,036 | 650,660 | 650,000 |
| Blog-blog links | 435,675 | 1,893,187 | 1,451,069 |
| Unique links | 245,840 | 648,566 | 1,158,803 |
| Average degree | 5.47 | 5.73 | 4.47 |
| Indegree distribution | -2.07 | -2.0 | -1.71 |
| Outdegree distribution | -1.51 | -1.6 | -1.76 |
| Degree correlation | 0.056 | 0.002 | 0.10 |
| Diameter | 14 | 12 | 6 |
| Largest WCC size | 96,806 | 263,515 | 617,044 |
| Largest SCC size | 4,787 | 4,614 | 72,303 |
| Clustering coefficients | 0.04429 | 0.0235 | 0.0242 |
| Percent Reciprocity | 3.03 | 0.6838 | 0.6902 |

Table 3: Comparison of blog network properties of datasets and simulation, Parameters: rR = 0.15, rW = 0.35, pD = 0.10, g = 1.06

**Conclusion.** We described a generative model for blog and post graphs that is based on the interactions among the bloggers and uses preferential and uniform random attachment techniques. We analyzed the key resulting structural properties, including degree distributions, degree correlations, reciprocity, average degree and clustering coefficient, and showed a power law distribution of connected components, posts per blog as observed in (Leskovec *et al.* 2007). The simulated Blogosphere will be useful to the research community for data generation, analysis and extrapolation.

**Acknowledgements.** Partial support was provided by NSF awards ITR-IIS-0326460 and ITR-IDM-0219649 and funding from IBM.

| Post network | ICWSM | WWE | Simulation |
|---|---|---|---|
| Total posts | 1,035,361 | 1,527,348 | 1,380,341 |
| Post-post links | 1,354,610 | 1,863,979 | 1,451,069 |
| Unique links | 458,950 | 1,195,072 | 1,442,525 |
| Avg post outlinks | 1.30 | 1.22 | 1.051 |
| Average degree | 2.62 | 2.44 | 2.10 |
| Indegree distribution | -1.26 | -2.6 | -2.54 |
| Outdegree distribution | -1.03 | -2.04 | -2.04 |
| Degree correlation | -0.113 | -0.035 | -0.006 |
| Diameter | 20 | 24 | 12 |
| Largest WCC size | 134,883 | 262,919 | 1,068,755 |
| Largest SCC size | 14 | 13 | 3 |
| Clustering coefficients | 0.0026 | 0.00135 | 0.00011 |
| Percent Reciprocity | 0.029 | 0.021 | 0.01 |

Table 4: Comparison of post network properties of datasets and simulation

## References

Chung, F., and Lu, L. 2006. *Complex Graphs and Networks (Cbms Regional Conference Series in Mathematics)*. Boston, MA, USA: American Mathematical Society.

Karandikar, A.; Java, A.; Joshi, A.; Finin, T.; Yesha, Y.; and Yesha, Y. 2007. Second space: a generative model for the blogosphere. Technical report, Univ. Maryland, Baltimore County.

Kolari, P.; Java, A.; Finin, T.; Oates, T.; and Joshi, A. 2006. Detecting spam blogs: A machine learning approach. AAAI 2006 - AI on the Web.

Lenhart, A., and Fox, S. 2006. Bloggers: A portrait of the internet's new storytellers. http://www.pewinternet.org/. Pew Internet and American Life Project Survey.

Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N.; and Hurst, M. 2007. Cascading behavior in large blog graphs. In *SIAM Int. Conf. on Data Mining*.

Pennock, D. M.; Flake, G. W.; Lawrence, S.; Glover, E. J.; and Giles, C. L. 2002. Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences* 99(8):5207–5211.

Vazquez, A. 2003. Growing networks with local rules: preferential attachment, clustering hierarchy and degree correlations. *Physical Review E* 67:056104.