Tracing the Birth of an OSN: Social Graph and Profile Analysis in Google+

Doris Schiöberg

doris@net.t-labs.tu-berlin.de

Fabian Schneider (NEC), Harald Schiöberg (TUB/T-Labs),
Stefan Schmid (TUB/T-Labs), Steve Uhlig (QMUL),
Anja Feldmann (TUB/T-Labs)

3.7.2012

Introduction

- Google+:
  - Google's social platform: social search
  - Profiles
  - Circles concept allows unidirectional links between users
- Chance to observe the early stage of an OSN
- OSN at all?

Agenda

- Introduction
- *Crawling Methodology and Data Sets*
- The G+ Graph
- Profile Data Analysis
- Conclusion

Crawling Methodology

1 Python Script, 1 Server, 1 Day, Full Graph

- Step 1
  - robots.txt → profiles-sitemap.xml → sitemap-0*.txt
  - Sitemap-*.txt:
    - Up to 5000 URLs of Google/Google+ profiles.
    - Each URL contains a user ID.
- Step 2
  - Download JSON objects describing the in/out-bound edges of a user Note: crawling limit per user is 10000
  - Compile graph, note unknown UIDs, Download data for unknown UIDs, Repeat if necessary ...

The Data Sets

- Google+ data:
  - Crawl over 6 weeks, 2/9 to 20/10, 16 data sets
  - One profile data set
  - Complete graph data + profiles from 20.10.2011
  - Some special properties:
    - Link Geo location correlations
    - Find the small islands

The Data Sets

- Twitter* and Flickr** data provided by Meeyoung Cha (KAIST)! Many thanks!

*Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K.: Measuring User Influence in Twitter: The Million Follower Fallacy. In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM) (2010).

**Cha, M., Mislove, A., and Gummadi, K. P. A: Measurement-driven Analysis of Information Propagation in the Flickr Social Network. In Proceedings of the 18th international conference on World wide web (WWW) (2009).

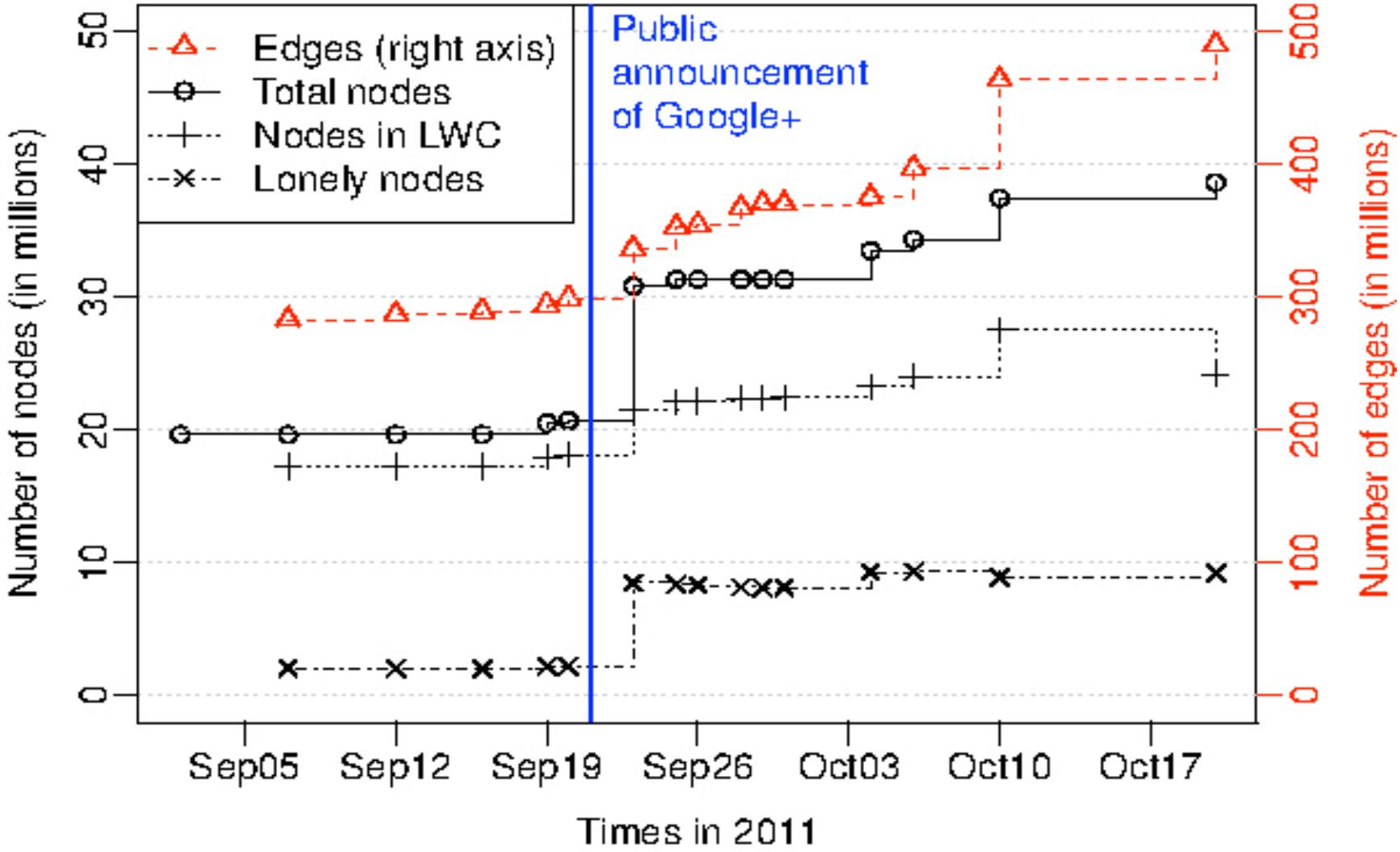
Agenda

- Introduction
- Crawling Methodology and Data Sets
- *The G+ Graph*
- Profile Data
- Conclusion

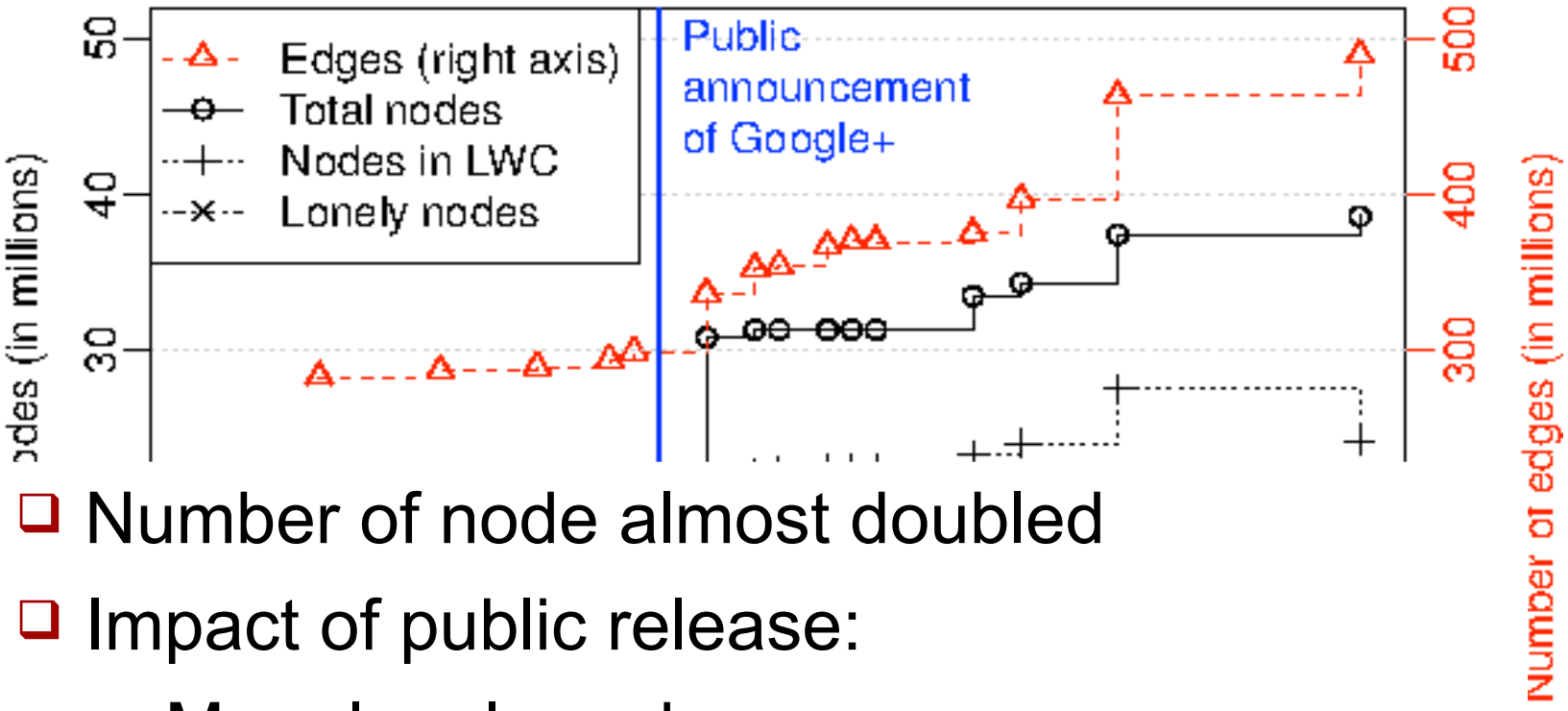
The Google+ Graph

- LWC: largest weak component
- Lonely node: user has no one in his/her circles

The Google+ Graph: Growth

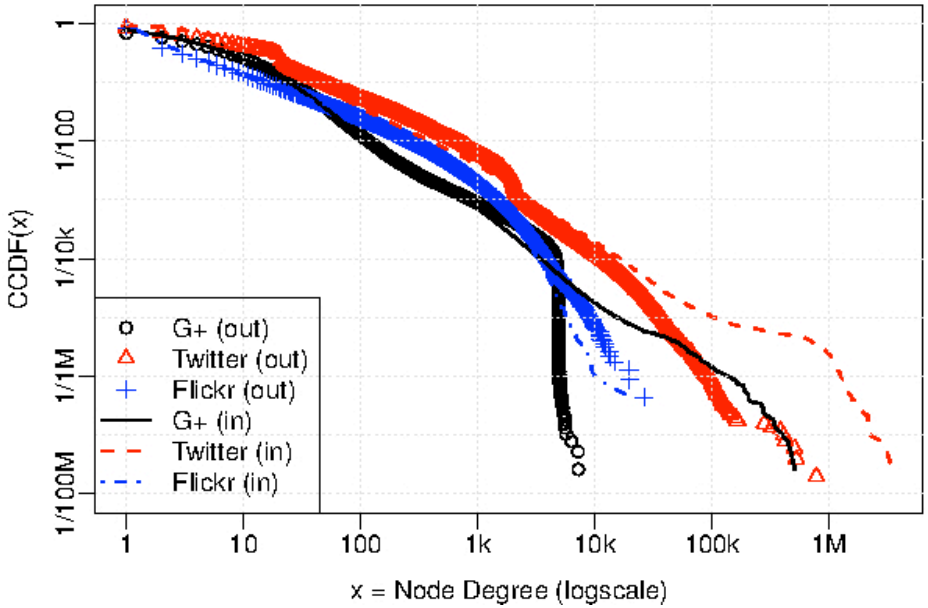


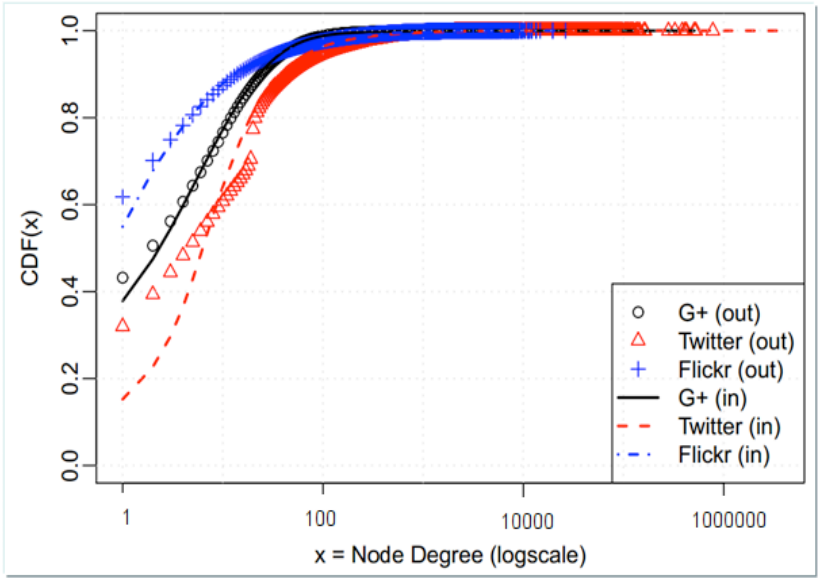
The Google+ Graph: Growth



- Number of node almost doubled
- Impact of public release:
  - More lonely nodes
  - More small islands
  - LWC relatively stable

The Google+ Graph: Degree





Closer to Flickr in the distribution body but closer to Twitter in the tail.

The Google+ Graph: Degree Correlation

Table 1. In/Out-degree correlation for OSNs.

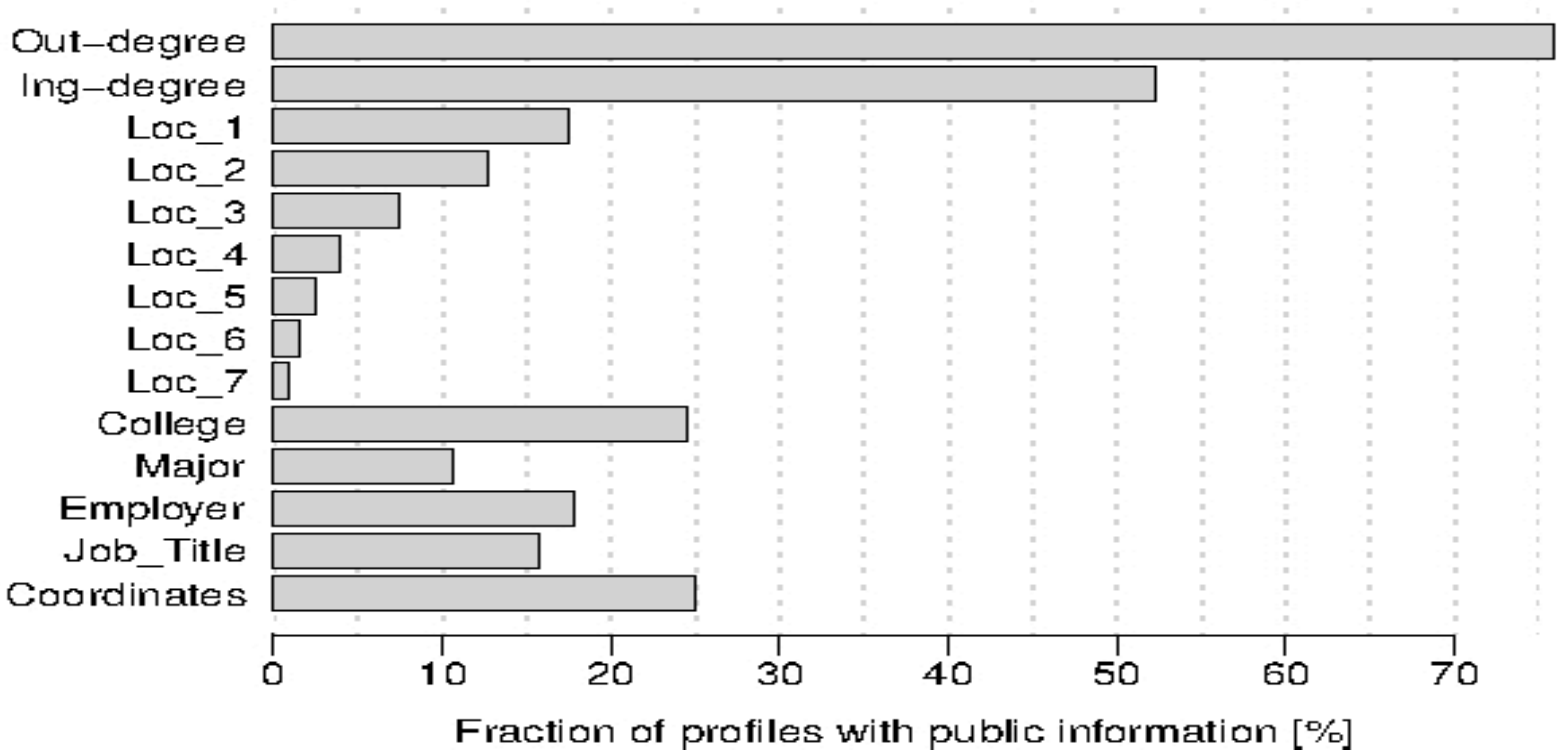
Graph	Nodes (in Mio.)	Degree Correlation
Google+	38.5	0.11606
Twitter	51.2	0.24532
Flickr	2.3	0.75584

Degree correlation coefficient is between Twitter and Flickr

Agenda

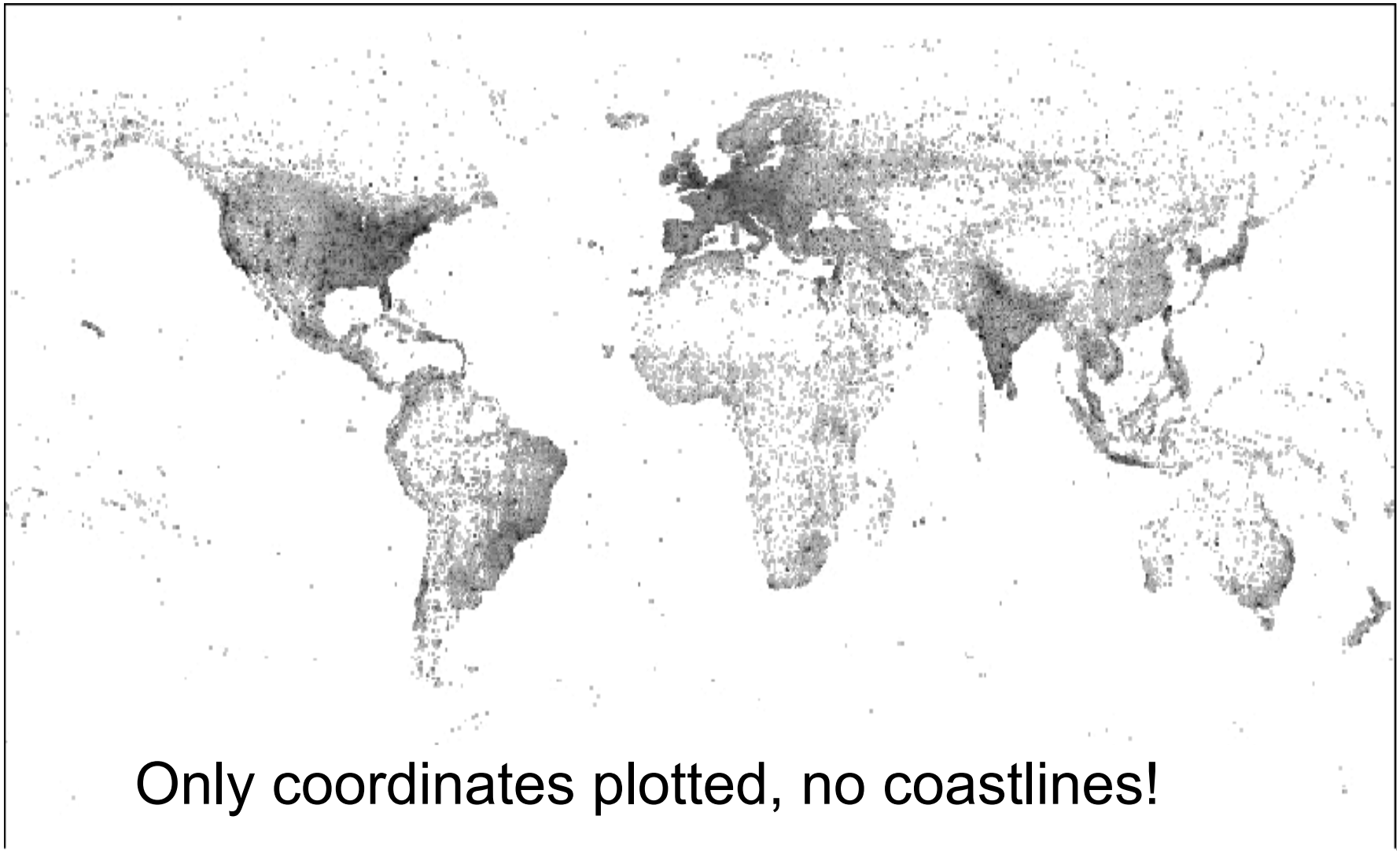
- Introduction
- Crawling Methodology and Data Sets
- The G+ Graph
- *Profile Data*
- Conclusion

Google+ Profiles



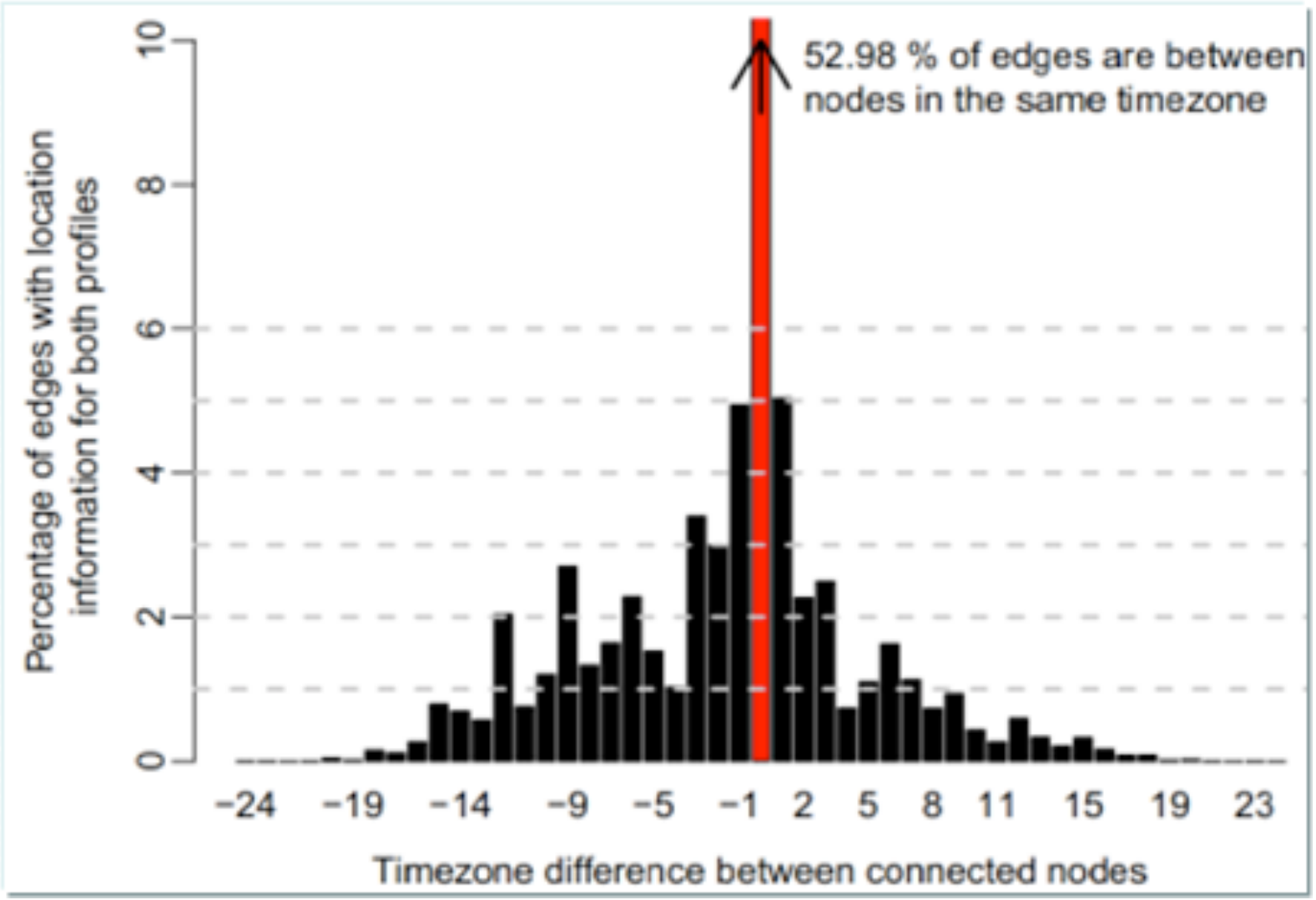
- Most (75%) show out-degree
- More info for only ~25% of the users available

Google+ Locations

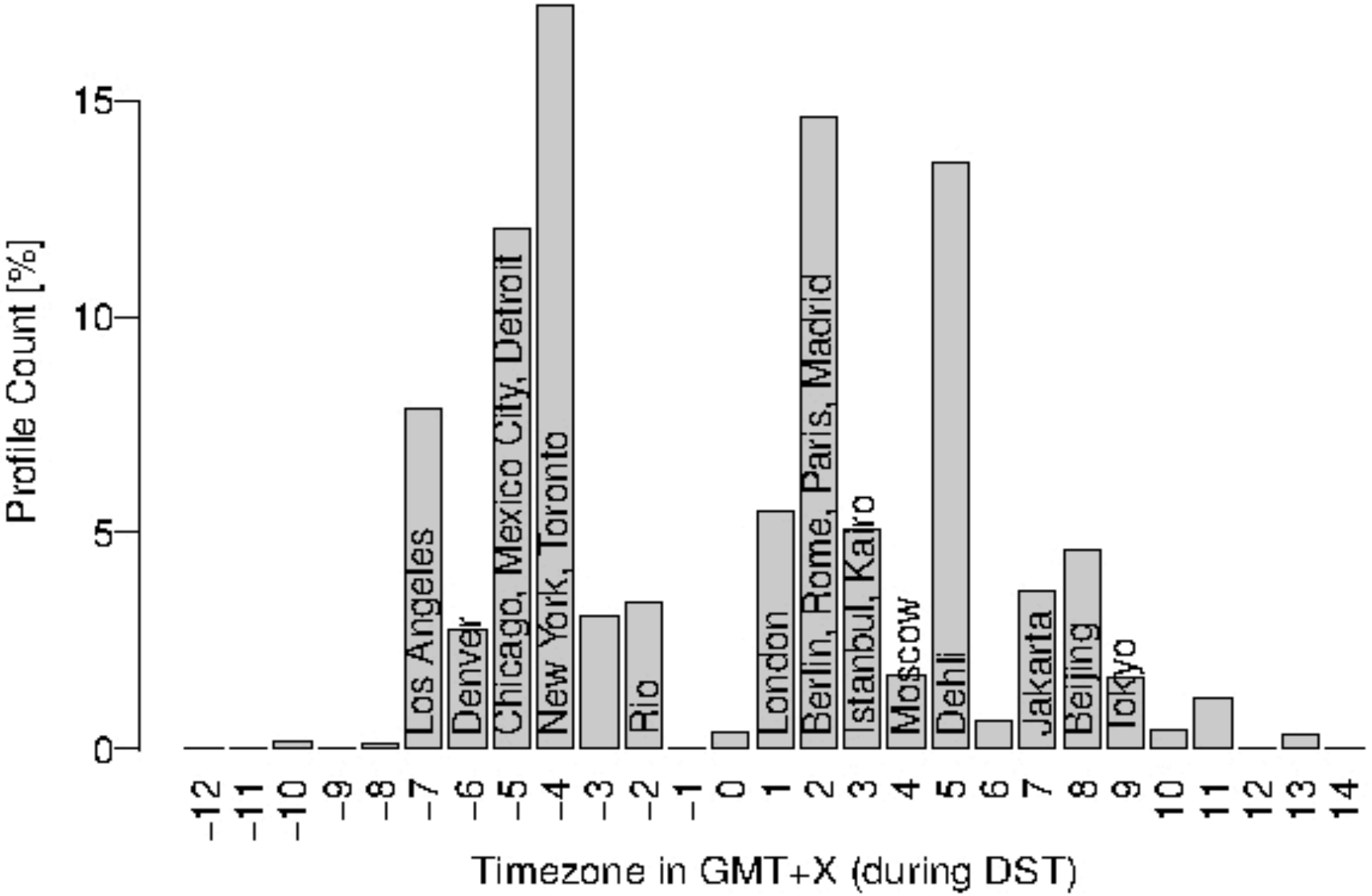


Only coordinates plotted, no coastlines!

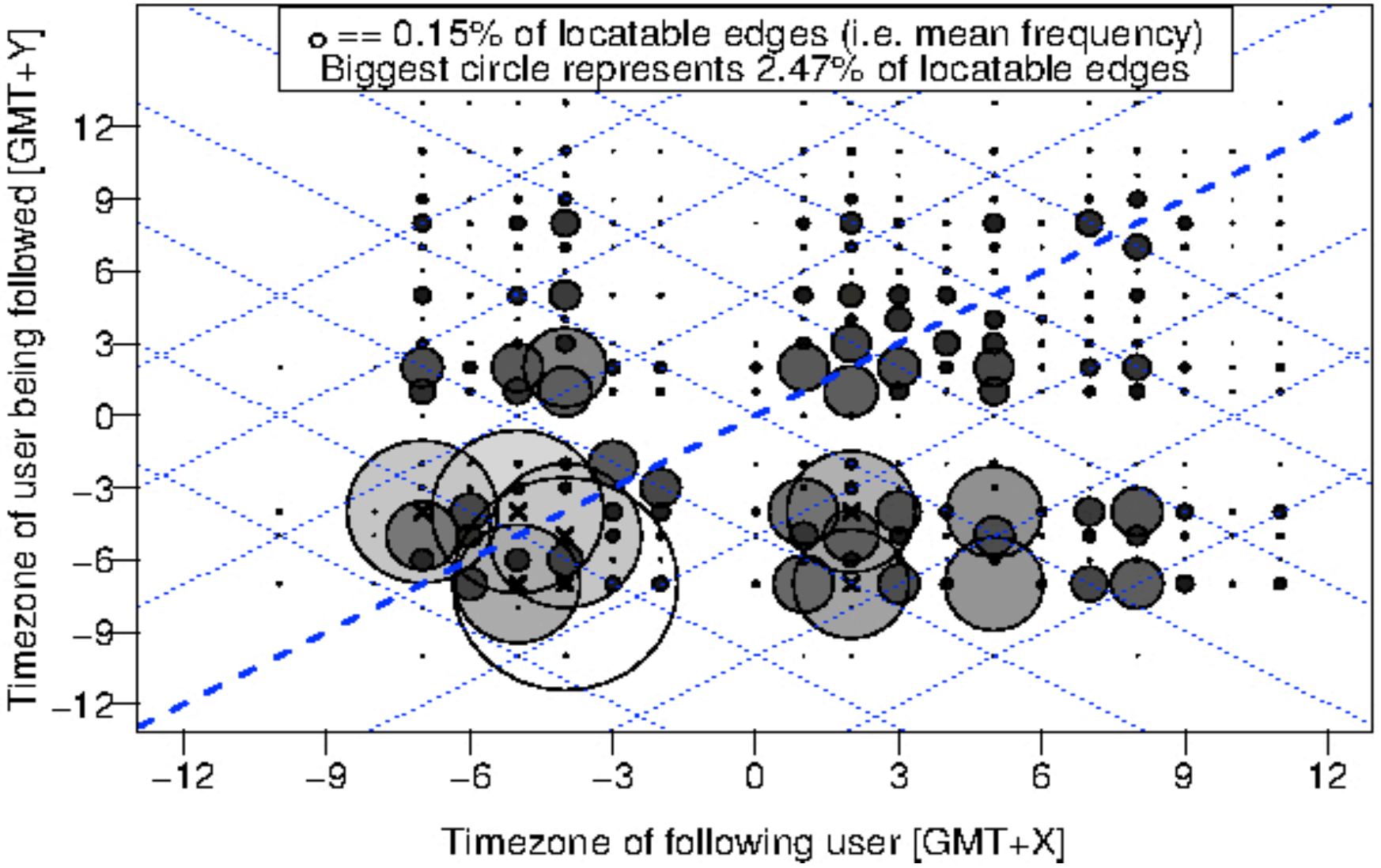
Google+ Distances



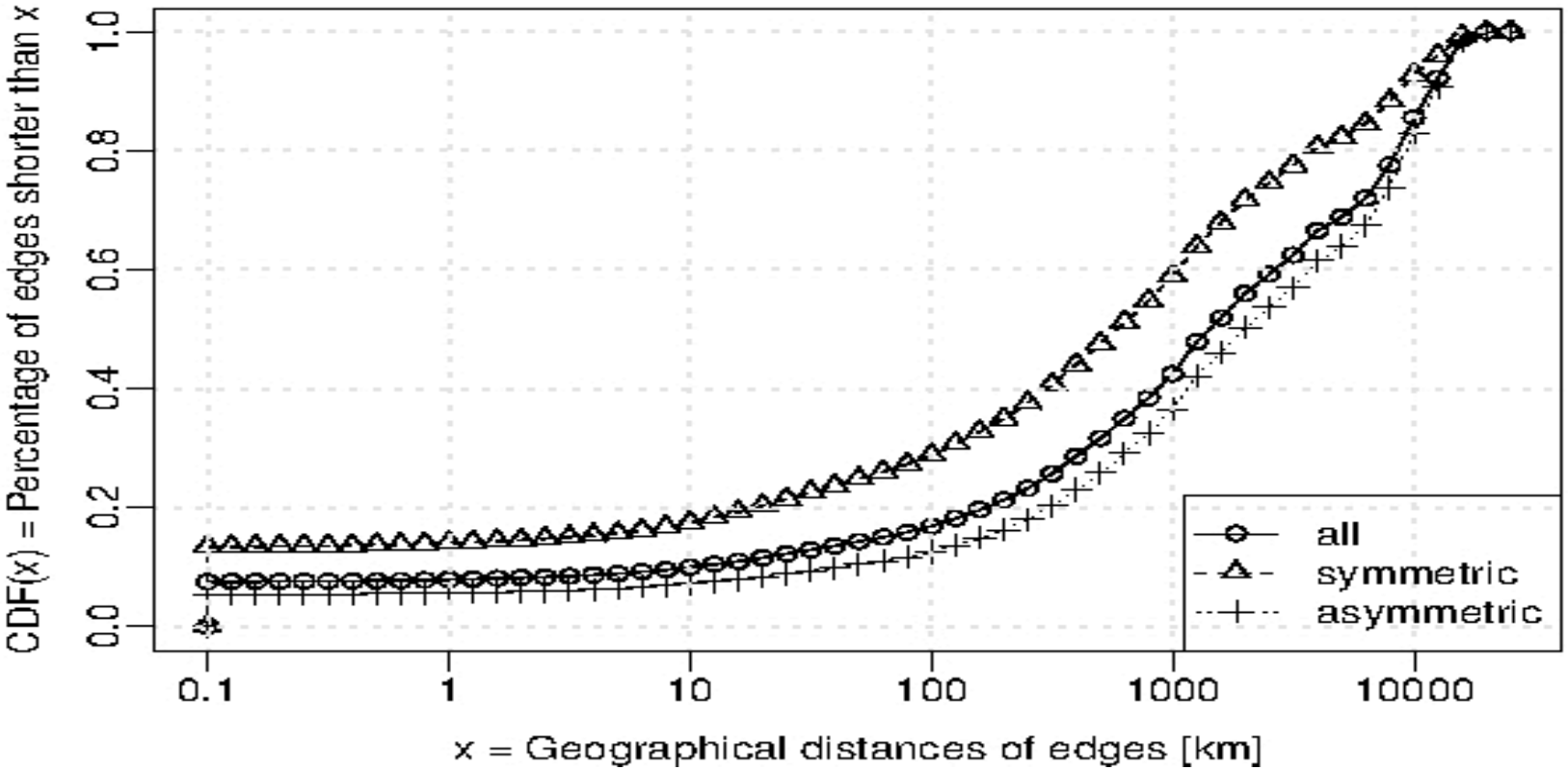
Google+ Timezones



Google+ Directions



Google+ Distances



- Symmetric links tend to be shorter than asymmtric (unidirectional) links

Agenda

- Introduction
- Crawling Methodology and Data Sets
- The G+ Graph
- Profile Data
- *Conclusion*

Summary

- Social Network Evolution:
  - Doubled in size over 6 weeks
  - Notable impact of public release
- Google+ Inhabitants:
  - Come from all over the world
  - 75% care about privacy
- Link ↔ Geo-location correlation:
  - >50% do not cross time-zones
  - East-west tendency for asymmetric links