How did Wikipedia become navigable?

Gergely Odor, Emanuele Bugliarello, Robert West

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

dlab — Data Science Lab

INTRODUCTION

The primary purpose of the hyperlinks on Wikipedia is to "aid navigation and understanding". While any user can try to ensure this property locally, no user can have a global view of Wikipedia. On the other hand, being well-navigable is a complicated global propery, that is seamingly still satisfied for Wikipedia [1]. We are interested in studying this process.

For measuring navigability we simulate a greedy search algorithm that tries to find the target using only local information and a network embedding (which approximates the mental map that people use for navigating Wikipedia). Since this notion of navigability has been well studied theoretically, we also revisit the question of modelling the evolution of the Wikipedia hyperlink network.

## THE GROWTH OF WIKIPEDIA

Models

1. Preferential attachment [2]

 - Let R1+R2+R3=1. Each new vertex is added to M others
 - Wp. R1, the edge leaves the new vertex pointing to an existing one chosen with prob. proportional to its in degree
 - Wp. R2 , the edge points to the new vertex, and the source vertex is chosen with prob. proportional to its out degree.
 - Wp. R3, the edge is added between existing vertices

2. Community guided attachment [3]

$$f(g) = c^{-h}$$

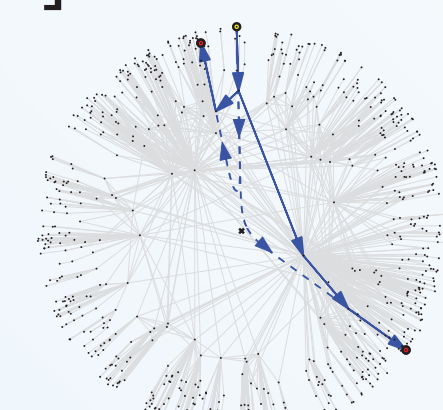h: tree distance
f: difficulty function

http://mathworld.wolfram.com/CompleteTernaryTree.html
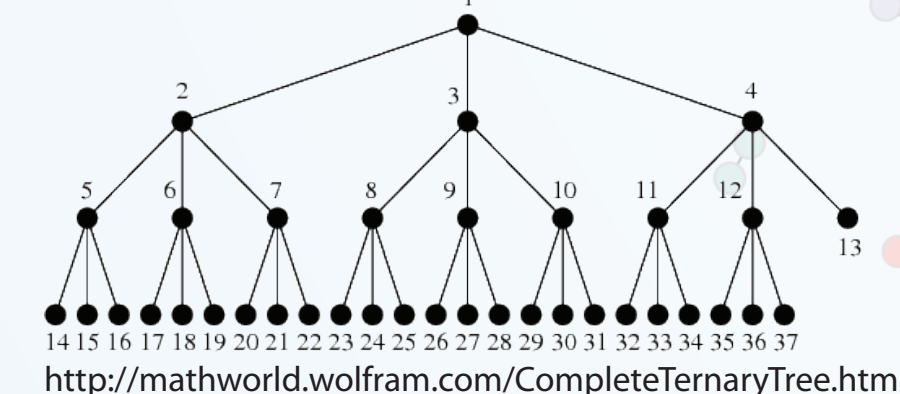
"When c<b, the average node degree is $n^{1-log_b(c)}$ and the in-degrees follow a Zipf distribution with exponent $0.5log_b(c)$"
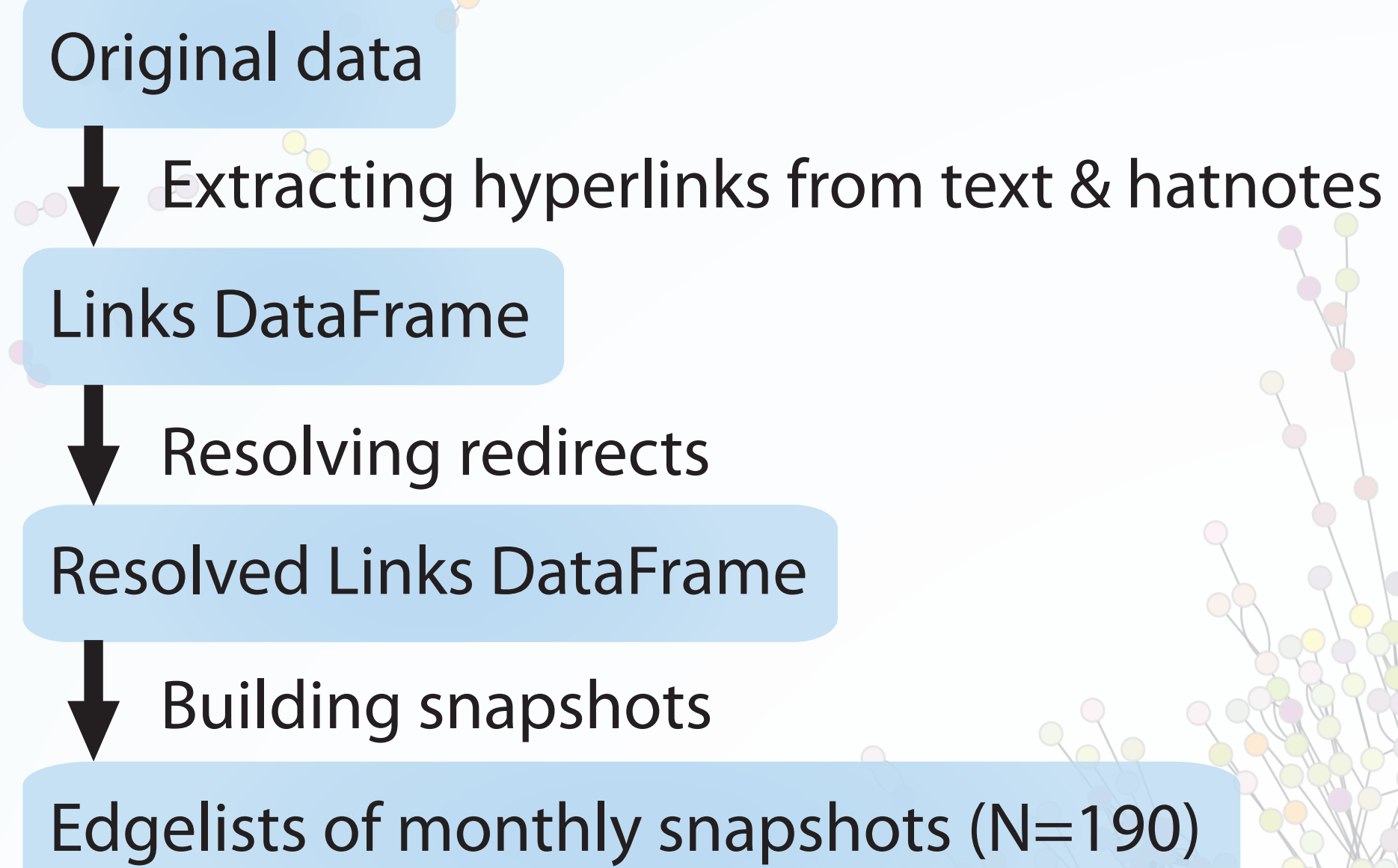
3. Popularity-Similarity Optimization [4]

 - Sample points over the hyperbolic disk
 - Connect points using hyperbolic distance
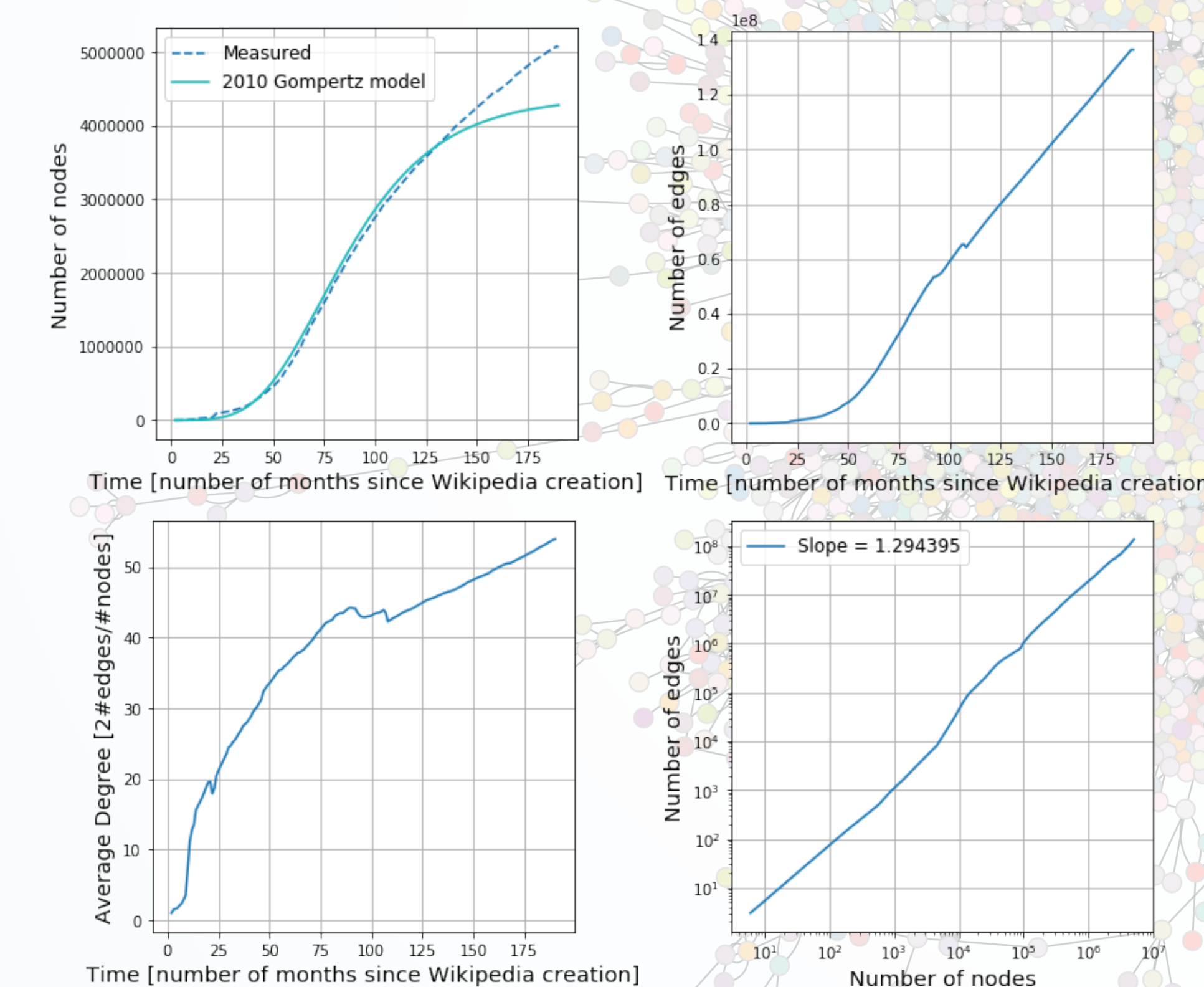
Hidden hyperbolic geometry has been shown to be a good model for scale-free networks with hierarchical organization [5]

Data preparation

Original data

↓ Extracting hyperlinks from text & hatnotes

Links DataFrame

↓ Resolving redirects

Resolved Links DataFrame

↓ Building snapshots

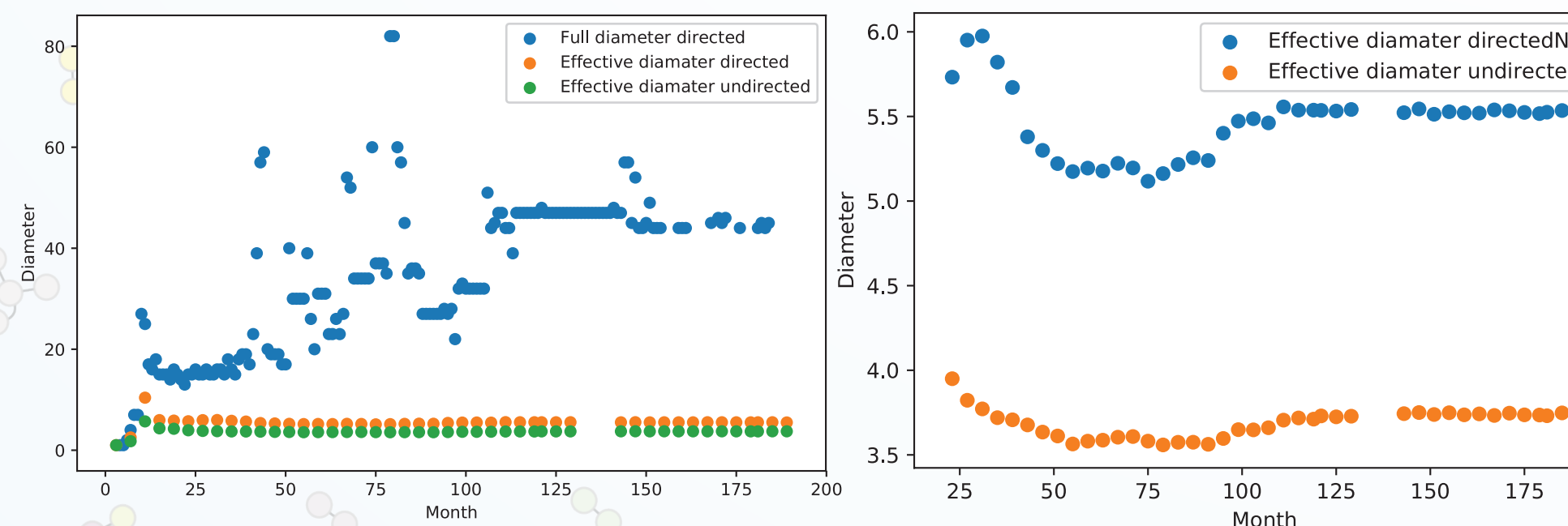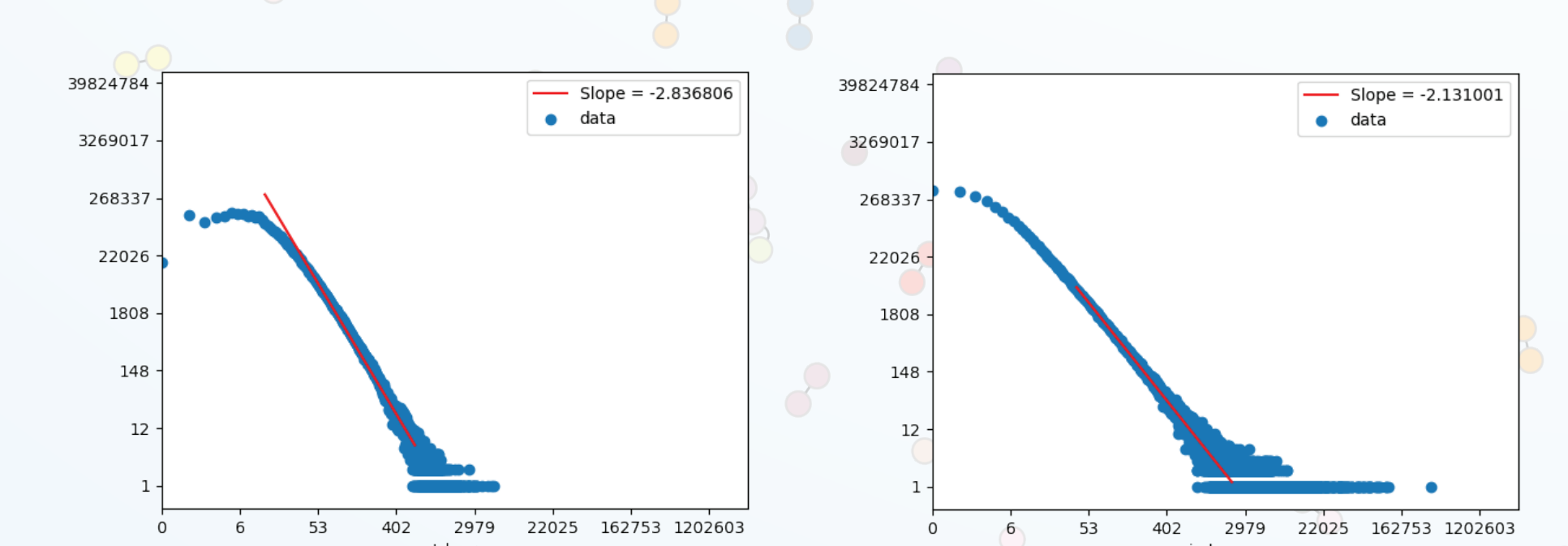Edgelists of monthly snapshots (N=190)

Results

Network growth over time
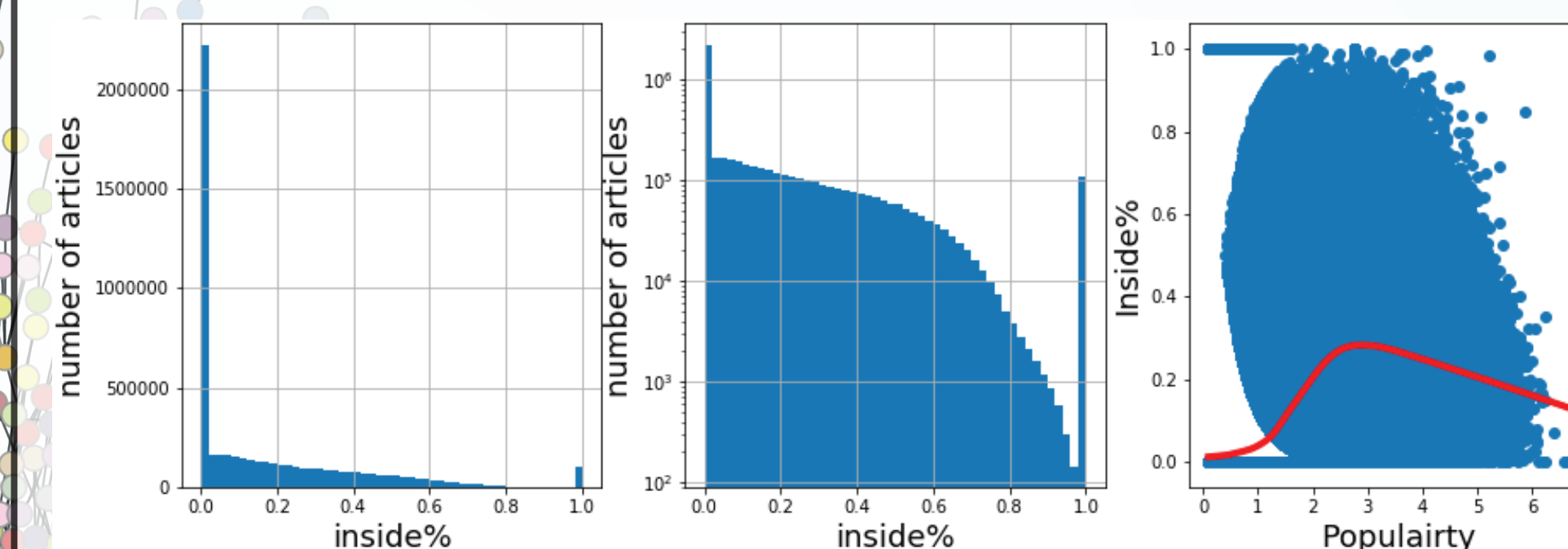

Diameter over time

Degree distribution

NAVIGABILITY

How do Wikipedia guidelines encourage navigability?

1. The guidelines say links should aid naviagation:
 - "Provide links that aid navigation and understanding, but avoid cluttering the page with obvious, redundant and useless links." [6]
Note: most traffic comes from outside

Inside% is the percentage of visits coming through an internal link

2. Reflects the "true" concept network, which might already be navigable
3. Efforts to make every article reachable
 - Red links
 - WikiProject_Orphanage

How do networks in general become naviagble?

1. Rewiring after unsuccesful navigation [7]
 - Unlikely on Wikipedia
2. Local clustering and long range connections [8]
 - More likely if we consider CGA or PSO

NOTE: to the best of our knowledge, all existing theoretical models consider navigability on an undirected network

Challenges of measuring navigability

1. Choice of embedding
 - Using article text only (e.g. LDA) or using network information ?

2. Choice of navigation algorithm
 - Greedy routing or other?
 - Choosing the soruce and the target

3. Interpreting results
 - Navigability = success probability does not tend to zero ? [8]
 - Navigability = Logarithmic path length ? [10]

Experimental results

Greedy routing
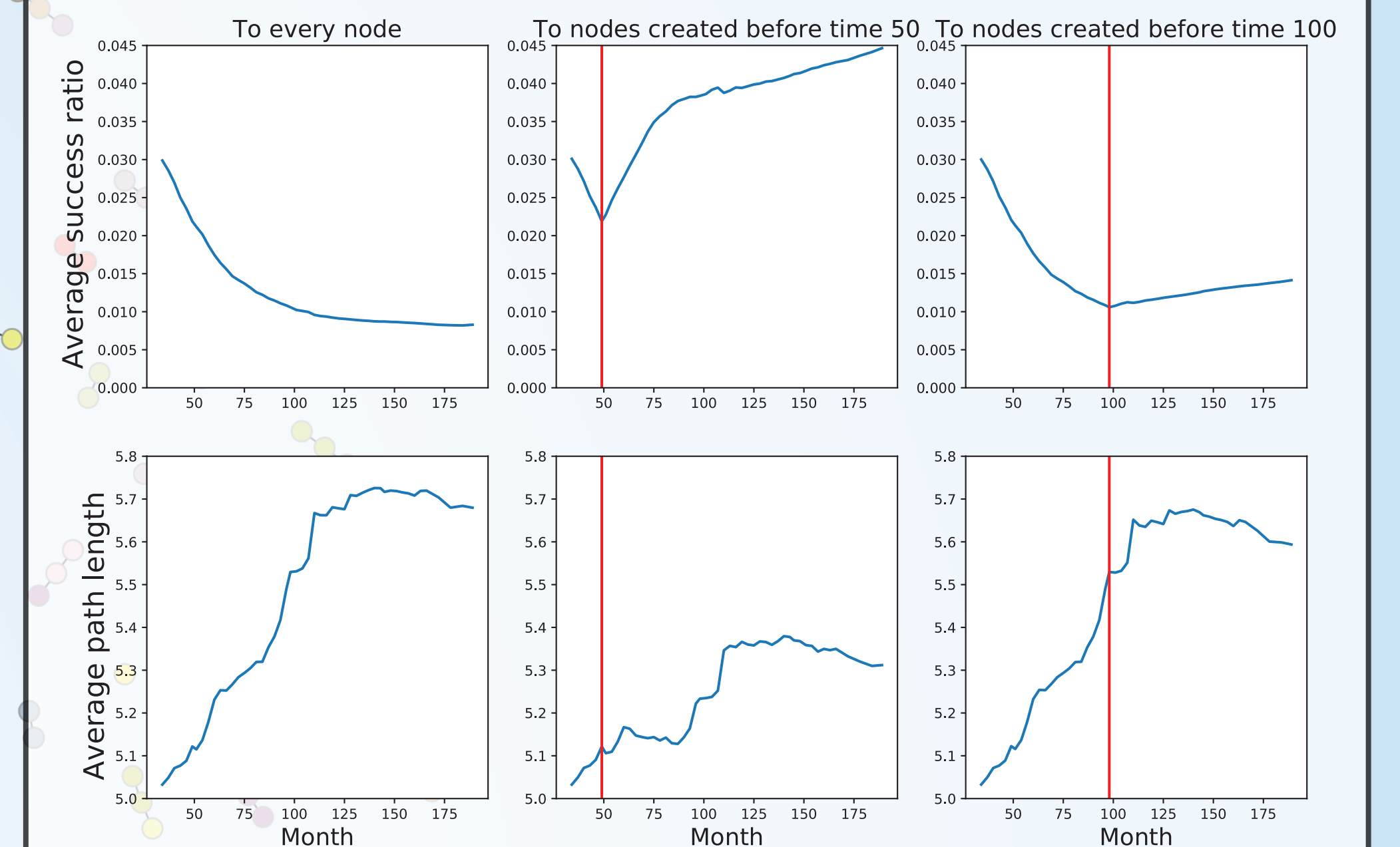
Depth-first search

FUTURE DIRECTIONS

1. Understand which links are the most important for navigation in Wikipedia

2. Study the effect of different in- and outdegree distributions on navigability in theoretical models

3. Propose a new network model for Wikipedia

REFERENCES

[1] R. West and J. Leskovec, "Human Wayfinding in Information Networks," in Proceedings of the 21st International Conference on World Wide Web, ser. WWW '12. New York, NY, USA: ACM, 2012, pp. 619–628.

[2] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, "Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia," Phys. Rev. E, vol. 74, no. 3, p. 036116, Sep. 2006.

[3] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph Evolu- tion: Densification and Shrinking Diameters," ACM Trans. Knowl. Discov. Data, vol. 1, no. 1, Mar. 2007.

[4] F. Papadopoulos, M. Kitsak, M. . Serrano, M. Bogu, and D. Krioukov, "Popularity versus similarity in growing net- works," Nature, vol. 489, no. 7417, p. 537, Sep. 2012.

[5] Krioukov, Dmitri, et al. "Hyperbolic geometry of complex networks." Physical Review E 82.3 (2010): 036106.

[6] Wikipedia. 2018. List of guidelines. (2018). https://en.wikipedia.org/wiki/ Wikipedia:List_of_guidelines

[7] A. Clauset and C. Moore, "How Do Networks Become Navigable?" Oct. 2003.

[8] Boguna, Marian, Dmitri Krioukov, and Kimberly C. Claffy. "Navigability of complex networks." Nature Physics 5.1 (2009): 74.

[9] W. Huang, S. Chen, and W. Wang, "Navigation in spatial networks: A survey," Physica A: Statistical Mechanics and its Applications, vol. 393, no. Supplement C, pp. 132–154, Jan. 2014.

[10] Kleinberg, "Navigation in a small world," Nature, vol. 406, no. 6798, p. 845, Aug. 2000.]