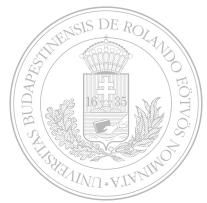
Twitter : Navigability

Reproduced report

Balázs Pál

Original author : Gábor Németh
Original supervisor : Eszter Bokányi
Eötvös Loránd University

Data Science Laboratory, December 10, 2020



Motivations and goals of topic

- Gábor's work was based on an already existing article on the problem[1]

[1] Szüle, J., Kondor, D., Dobos, L., Csabai, I., & Vattay, G. (2014). Lost in the City: Revisiting Milgram's Experiment in the Age of Social Networks. PloS one, 9(11), e111973.

Motivations and goals of topic

- Gábor's work was based on an already existing article on the problem[1]
- Detect indications of the small-world phenomenon in US Twitter user data
  - Mostly **short path lengths** in the graph/network of user nodes
  - "Connections" are simply defined as users following other users
  - Twitter has an extra information: *Geographical data*

[1] Szüle, J., Kondor, D., Dobos, L., Csabai, I., & Vattay, G. (2014). Lost in the City: Revisiting Milgram's Experiment in the Age of Social Networks. PloS one, 9(11), e111973.

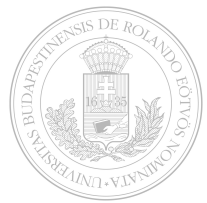
Motivations and goals of topic

- Gábor's work was based on an already existing article on the problem[1]
- Detect indications of the small-world phenomenon in US Twitter user data
  - Mostly **short path lengths** in the graph/network of user nodes
  - "Connections" are simply defined as users following other users
  - Twitter has an extra information: *Geographical data*
- Explore the **navigability** of Twitter users' network
  - A network is navigable, if some decentralized algorithm can find the path between two arbitrary nodes
  - The decentralized algorithm of choice for this problem is the **Greedy algorithm**

[1] Szüle, J., Kondor, D., Dobos, L., Csabai, I., & Vattay, G. (2014). Lost in the City: Revisiting Milgram's Experiment in the Age of Social Networks. PloS one, 9(11), e111973.

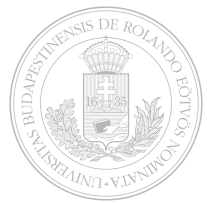
Dataset and preprocessing

- Pretty large dataset originally (100+ million edges)
  - Contains user ID's, edges and geo-tags
  - Requires a lot of computational time to preprocess
  - Requires a lot of RAM to even load it, and we also need to execute the algorithm using the data too



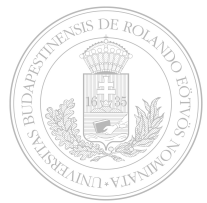
Dataset and preprocessing

- Pretty large dataset originally (100+ million edges)
  - Contains user ID's, edges and geo-tags
  - Requires a lot of computational time to preprocess
  - Requires a lot of RAM to even load it, and we also need to execute the algorithm using the data too
- Dataset I've worked with was just a subset of nodes of the original network
  - Nodes were chosen around Missouri area and only the giant component was kept: 39501 nodes and 193299 edges
  - Positional coordinates (latitudes-longitudes) were also included



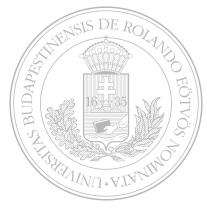
Dataset and preprocessing

- Pretty large dataset originally (100+ million edges)
  - Contains user ID's, edges and geo-tags
  - Requires a lot of computational time to preprocess
  - Requires a lot of RAM to even load it, and we also need to execute the algorithm using the data too
- Dataset I've worked with was just a subset of nodes of the original network
  - Nodes were chosen around Missouri area and only the giant component was kept: 39501 nodes and 193299 edges
  - Positional coordinates (latitudes-longitudes) were also included
- Some other preprocessing steps were also made
  - Creating a neighbour list for every nodes
  - Creating a dataset with user ID's and tweet locations
  - Selecting the edges in the examined smaller area



My plans

- Numerous options for "to show something new"
  - Preprocess the original dataset and create new, smaller datasets for other regions (eg. for other cities)
  - Implement and experiment with other decentralized algorithms
  - Create new visualizations
- My actual plans and work
  - Reworked the current algorithms into a more compact and optimized form
  - Omitted unused or unnecessary functions
  - Created several functions for recurrent tasks
  - Written informative comments and docstrings for better usability
  - Completely reworked figures
    - Reworked style and outlook
    - Added new relevant informations (eg. fits/KDEs)



Results - Degree distribution

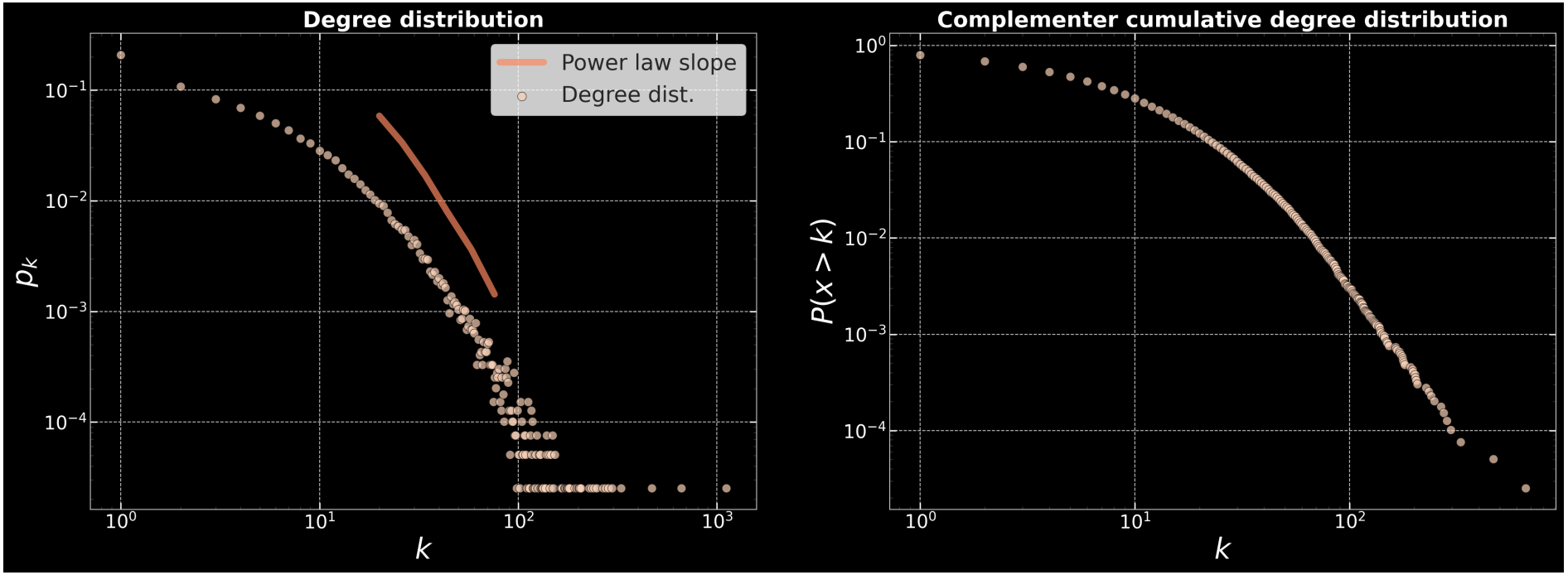


Figure 1: Degree distribution (left) and complementer cumulative degree distribution (right) of the selected nodes around Missouri. The degree distribution follows the power law, $P(k) \sim k^{-\gamma}$, which indicates the selected sub-graph is could be a scale-free network between $20 < k < 100$. In this interval the exponent $\gamma = 2.604 \pm 0.023$. The clustering coefficient $C = 0.172$ [2].

[2] For a much larger dataset in the original article these values were $\gamma = 2.60 \pm 0.01$ and $C = 0.14$.

Results - Distance distribution

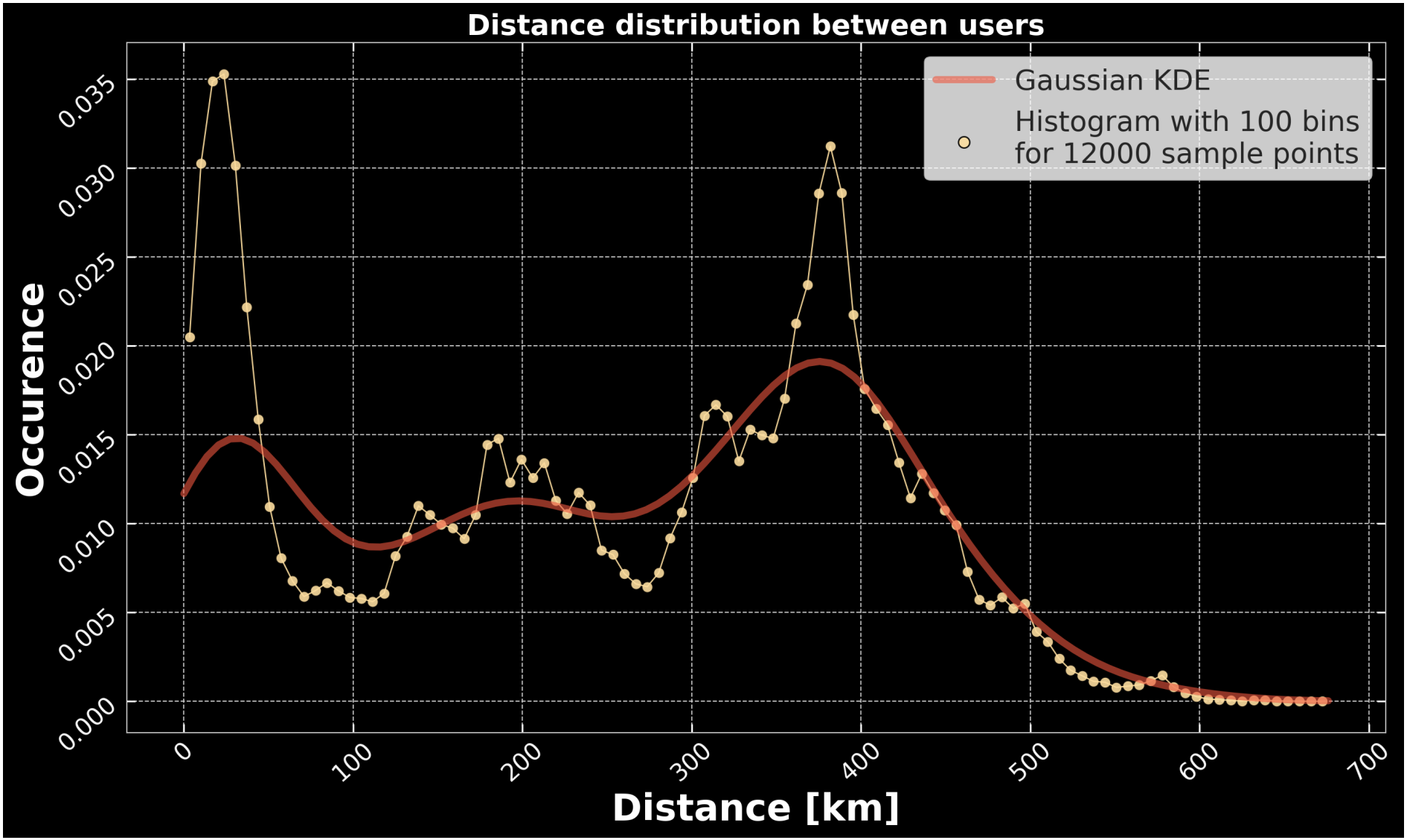


Figure 2: The geographical distance distribution between 12000 randomly selected nodes of the network.

Results - Greedy algorithm

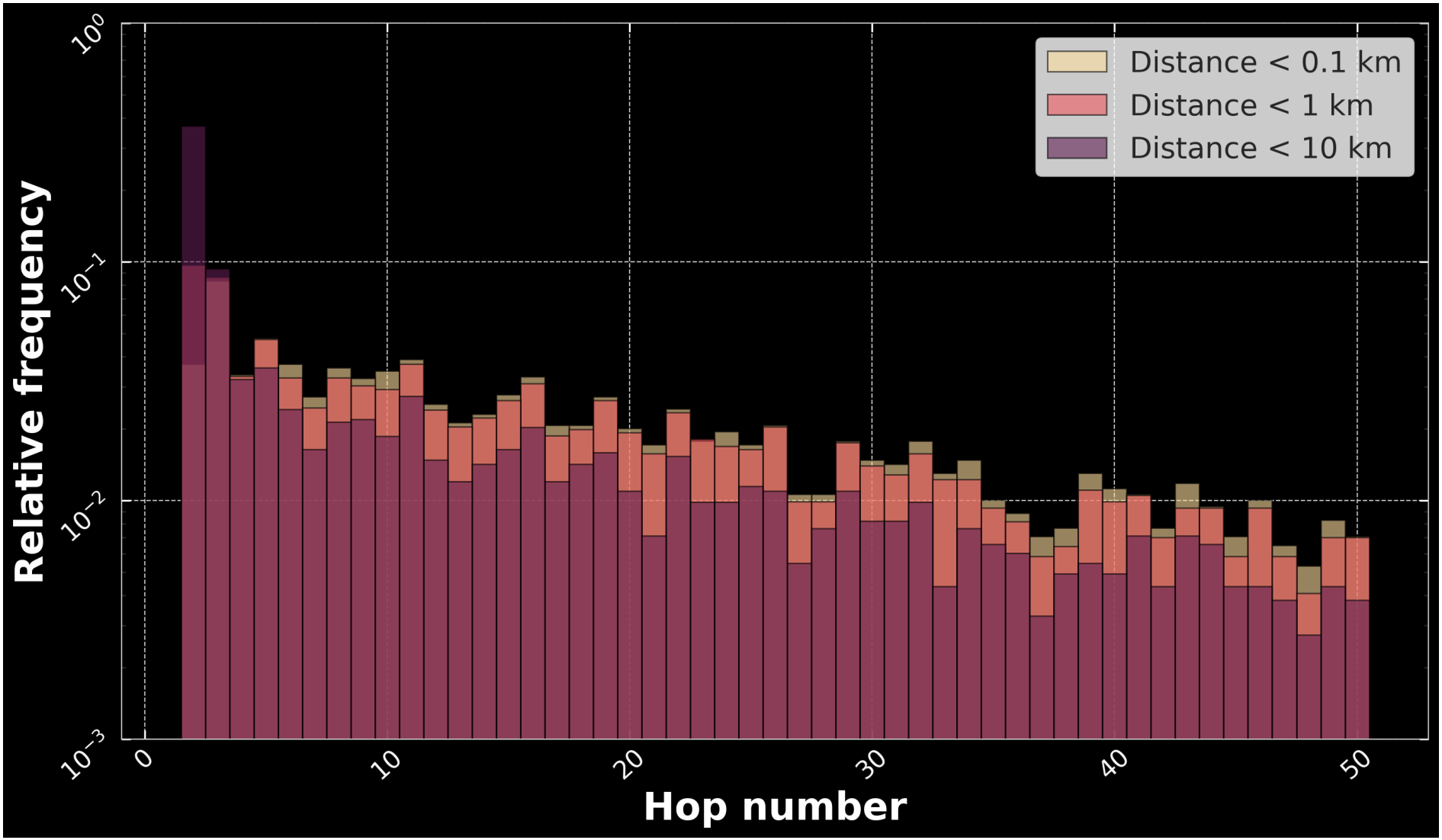


Figure 3: Number of successful greedy searches with different distance thresholds and an upper limit for the number of hops.