Cover letter

May 31, 2019, Copenhagen, Denmark

Application for Research Scientist position at Wikimedia Foundation

I would like to be considered for the position as Research Scientist at the Wikimedia Foundation. I am currently a PhD student at the University of Copenhagen (UCPH) pursuing a doctorate in social data science at the Center for Social Data Science (SODAS). I am also affiliated with the Department of Applied Mathematics and Computer Science at the Technical University of Denmark (DTU Compute). My background is in Physics and Computer Science and I obtain my doctorate in August, this year. Below I will describe what I do and how it applies to this position; I will motivate why I think I am a good candidate, and finally propose a feasible remote working scenario.

My research is on mapping large scale patterns in human behavior. It relies on a first-of-its-kind social data experiment which ran from 2013 to 2016, where ~1000 undergraduate university students (~80% of a tight-knit population) were given free personal smartphones that continuously donated high-resolution behavior and mobility data to a protected database [1]. **In one project**, I built a community detection algorithm (in C++ with a wrapper for Python) for temporal networks and applied it to a large dynamic face-to-face network reconstructed from Bluetooth signals, to find social cliques [2] (see Fig. 1 or Ref [3] for an interactive version). In a follow-up study, I applied the same method to brain fMRI data, to recover functional networks [4] (see Fig. 2). This line of research demonstrates that I can work practically with big complex network datasets at many levels of abstraction. Wikipedia is a vast multiplex and temporal network of information flow and user interaction, so I consider this skill highly relevant. **In my current project**, I work on human mobility using high-resolution location data from GPS and WiFi. The work aims at modeling mobility as a multi-scale phenomenon, and in an unsupervised manner using just raw mobility data, to infer a hierarchy of nested “containers” for mobility in physical space (e.g. continents, countries, cities, neighborhoods, etc., see Fig. 3). The mobility model uses Bayesian inference for labeling the scale of each trip, and in validation I use coding theory and as well as recurrent neural networks for measuring the information gain and increased location forecasting accuracy that comes from a good multi-scale descriptions of mobility. I imagine that my extensive experience with geographical data, and the tools I have learned in this context, can prove useful when working with data collected continuously from all around the world. **In a side project** I have worked with the Y-Combinator funded startup Peergrade on designing A/B testing experiments on their e-learning platform for understanding user behavior [5, 6]. Specifically, we designed multiple algorithms for distributing tasks to users and looked for the most efficient one with respect to various objectives. You request experience with large-scale experiments in online platforms in the job posting, and I think this project qualifies me in this regard because it taught me how to think deeply about user experience and write online algorithms that steer user behavior in real-time and interact with a live database. **I develop and maintain** a number of open source software packages (see my Github account [7]), most notably the recent *Netwulf* Python package that enables JavaScript/d3 powered interactive network visualization directly from the Python console [8]. I am strongly committed to distributing and/or publishing codes I produce that are general enough to be used elsewhere. Many of my projects (*Netwulf, Infostop, py_pcha)* are readily installable with the Python package manager *pip*.

I have an extensive teaching and speaking record. Over the past five years I have co-taught three different Master’s level courses and independently taught two Bachelor’s level courses on topics relating to Data Science and Machine Learning. Specifically, I have taught linear algebra, statistics, probability theory, various machine learning methods, neural networks, network science, text processing, MapReduce, and ethics and law around Big Data. I have supervised a large number of exam projects and Master’s thesis projects. I have given many talks at different conferences, workshops and labs, and co-organized workshops and events, most recently the Copenhagen chapter of the popular Europe-wide event phenomenon *DataBeers*. My teaching (at least once per week) and frequent public speaking have earned me the ability to communicate effectively, and confidently speak in front of large audiences. I have, furthermore, established a strong presence within the complex systems research community on Twitter. Taken together, these qualities would undoubtedly be an asset for increasing visibility and impact of work produced at Wikipedia.

I have received a number of awards during and before my PhD work. Science Magazine awarded me first place in their inaugural “Data Stories” competition in 2016 [9] for the data visualization hosted at Ref [3]. I received a communications award at the 2018 NetSci conference in Paris, as well as a “Best Paper" award at the 2017 International Conference on E-Learning in Florida. I have received funding from nine different sources to support academic visits to Israel, USA, Germany and Singapore.

I have two main motivations for seeking this position. (1) I love Wikipedia. I give a monthly donation to the Wikimedia Foundation because it is one of the few places on the Internet where I’m not exploited for data to serve companies that buy ad-space. I am a longtime follower of people like Roger McNamee, Jaron Lanier and Tristan Harris, and in line with their thinking I am enormously frustrated that most data science jobs serve this harmful business model directly or indirectly. I, therefore, see this job as an opportunity to do data science for social good and scientific research at the same time. (2) I have experience with Wikipedia data (the MediaWiki API is a core component in one of my courses), and I am well aware of what an awesome data mine of human behavioral patterns and motifs it is. For example, I think revision dynamics on articles describing controversial issues is exciting. In a time when debates over controversial issues are becoming increasingly heated, people naturally seek Wikipedia as a source of truth. Wikipedia is, therefore, an obvious target for vandalism and other types of abuse. Furthermore, significant differences in discourse across languages may arise (e.g. English vs. Russian version of the ‘Malaysia Airlines Flight 17’ article). Imaginably, lots of interesting new problems need solving in order to maintain Wikipedia as an objective source of knowledge.

I am applying as a remote candidate. My current academic environments at the Center for Social Data Science (SODAS) and DTU Compute are highly stimulating and productive places populated by social and natural scientists. We are frequently visited by top researchers from around the world, and recent publications from my peers are featured in publications like Nature Human Behavior, Nature Communications, PNAS and Scientific Reports [10-14]. Graduates have moved on to research positions at Google, Unicef, Northeastern University and MIT. If you are interested in continuing the conversation we might want to work out the specifics together, but let me propose the following scenario: *I could work for the Wikimedia Foundation full time and keep my current academic affiliations and desk spots*. This would allow me to build strong collaborations and conduct high-impact research. I could start Master’s and PhD students on projects leveraging your massive datasets to explore hypotheses at no expense to you. With high-speed internet I could check in frequently, and with a nearby airport that directly connects Copenhagen to San Fransisco I could fly in to headquarters from time to time.

Thanks for taking the time to read through my letter. I am looking forward to your response!

Ulf Aslak

References

[1] Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., & Lehmann, S. (2014). Measuring large-scale social networks with high resolution. PloS one, 9(4), e95978.

[2] Aslak, U., Rosvall, M., & Lehmann, S. (2018). Constrained information flows in temporal networks reveal intermittent communities. *Physical Review E*, *97*(6), 062312.

[3] http://ulfaslak.com/research/temporal_communities/

[4] Aslak, U., Nielsen, S. F. V., Mørup, M., & Lehmann, S. (2019). Temporally intermittent communities in brain fMRI correlation networks. *Journal of Applied Network Science*. Under review.

[5] Wind, D. K., Aslak, U., Jørgensen, R. M., Hansen, S. L., & Winther, O. (2017, October). Optimal allocation of reviewers for peer feedback. In European Conference on e-Learning (pp. 566-573). Academic Conferences International Limited.

[6] Wind, D. K., & Aslak, U. (2017, June). Quantifying Feedback: Insights Into Peer Assessment Data. In ICEL 2017-Proceedings of the 12th International Conference on e-Learning (p. 256). Academic Conferences and publishing limited.

[7] https://github.com/ulfaslak

[8] https://joss.theoj.org/papers/3a22c963a45dbddc8501a4b5ef4b2bf6

[9] https://www.sciencemag.org/projects/data-stories/winners/2016

[10] Lorenz-Spreen, P., Mønsted, B. M., Hövel, P., & Lehmann, S. (2019). Accelerating dynamics of collective attention. *Nature communications*, *10*(1), 1759.

[11] Sekara, V., Deville, P., Ahnert, S. E., Barabási, A. L., Sinatra, R., & Lehmann, S. (2018). The chaperone effect in scientific publishing. *Proceedings of the National Academy of Sciences*, *115*(50), 12603-12607.

[12] Alessandretti, L., Sapiezynski, P., Sekara, V., Lehmann, S., & Baronchelli, A. (2018). Evidence for a conserved quantity in human mobility. *Nature Human Behaviour*, *2*(7), 485.

[13] Sekara, V., Stopczynski, A., & Lehmann, S. (2016). Fundamental structures of dynamic social networks. *Proceedings of the national academy of sciences*, *113*(36), 9977-9982.

[14] Stopczynski, A., & Lehmann, S. (2018). How Physical Proximity Shapes Complex Social Networks. *Scientific reports*, *8*(1), 17722.

Figures

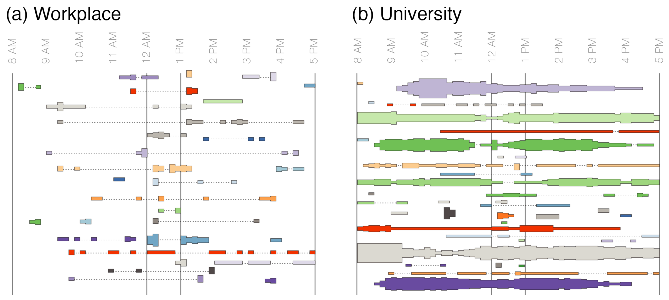

Figure 1. Temporal communities detected by Infomap with neighborhood flow coupling. Each horizontal track represents a community and its varying height represents the number of active nodes over time. (a) Partition of the workplace network. Height to scale with (b). (b) Partition of the university network. At its tallest point, the largest community (top purple, 10 am) has 22 active members.

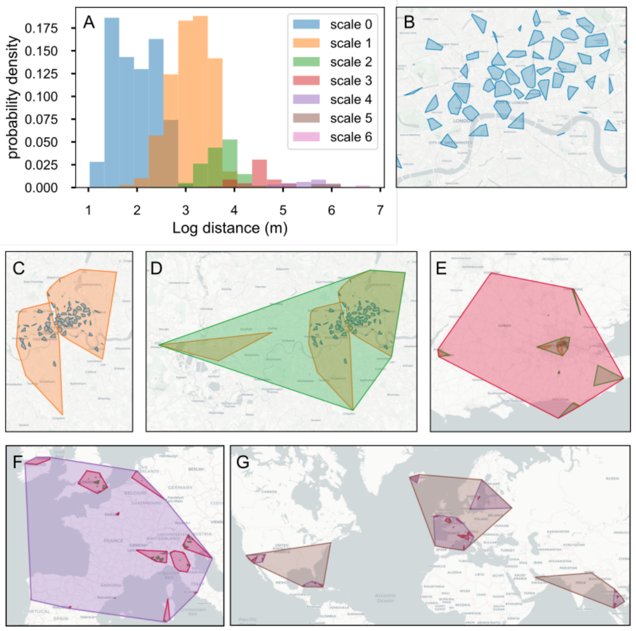

Figure 3. Example of spatial scales. The spatial scales recovered from the mobility trajectory of one of the authors. (A) Probability density of the logarithm in base 10 of the displacement length (in meters), for displacements occurring within containers at the same scale. Each colour represents a different scale. (B-G) Examples of containers at different scales, from lowest (B) to highest (G). Map tiles by Carto, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

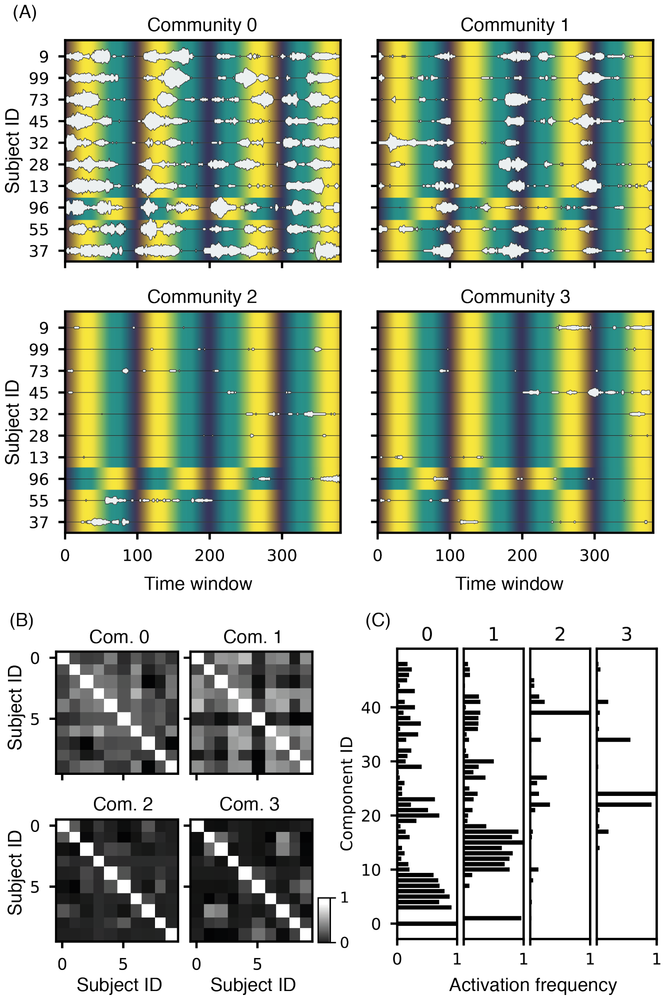

Figure 2. Temporal communities of 10 “best” performing subjects concatenated. (A) Four largest communities in the subject-concatenated network, reshaped to visualize one subject per row to enable comparison of temporal profiles. Notice that the task order deviates for subject 96. (B) Inter-subject correlation of temporal activity for the four largest communities. Mean correlations (not including diagonal of ones) are 0.33, 0.41, -0.024, and 0.022, respectively. (C) Component distributions. See 6 for component plots.