Computer Science Department
Direct phone: +45 7218 5284
E-mail: pagh@itu.dk

May 30, 2019

# Assessment of the PhD thesis of Piotr Wygocki, "On Nearest Neighbors"

Piotr Wygocki's thesis has two main parts:

1) Design and analysis of new randomized algorithms for approximate nearest neighbors that provide expected-time ("Las Vegas" type) guarantees, as opposed to the error probability guarantees provided by previous approaches in high-dimensional Euclidean space.

2) New models for information dissemination in social networks, designed to explain the appearance of "cascades" of information transmission (e.g. retweets).

**Assessment of part 1**

Approximate nearest neighbors in high dimensions is a fundamental problem that has been intensely studied in the last two decades. Until recently the fastest algorithms all came with a catch: They were "Monte Carlo" algorithms, i.e, used random coin tosses and would fail to provide a correct answer with some (small) probability. Pagh, and later Ahle, showed that this could be avoided in several settings such that the answer would always be correct, but the expected running time might be exceeded with some small probability ("Las Vegas" algorithms). However, they left open the question for the most studied setting of all, Euclidean space.

The thesis presents several algorithmic techinques that allow such error-free algorithms in Euclidean space. Some of the techniques are nice generalizations of methods used in Monte Carlo algorithms, for example:

- The thesis makes the fundamental observation that it is possible to efficiently reduce the d-dimensional problem to O(d/log n) problems in dimension O(log n), where n is the number of data vectors. Thus, up to a multiplicative overhead of O(d/log n) it suffices to consider the case of O(log n) dimensions, where the constant depends on the desired approximation factor.

- Random projections with bounded coordinates yield Las Vegas guarantees, and have properties similar to standard random projections when the approximation factor of interest is close to 1. In particular, the thesis studies anti-concentration results that show such projections tend have few points that are substantially closer after the projection compared to before.

The scalability of nearest neighbor algorithms is characterized by the exponents of the space usage and search time, which are usually written in the form n^alpha for some alpha. The exponents achieved in the thesis are close to the best known for approximation factor c close to 1, but worse for large values of c. We note that A. Wei, in a paper that was made public in July 2018 and appeared at SODA in January 2019, has

achieved stronger results that match the best known exponents in the Monte Carlo setting. This paper is not discussed in the thesis. Nevertheless, the techniques in the thesis are interesting: The are simpler than those of Wei, and many of them were published before Wei's paper, and may have contributed to arriving at the tight results.

**Assessment of part 2**

It is of high interest to understand viral content in social networks, also known as "cascades". The thesis discusses limitations of previously proposed models of information dissemination with respect to cascades, and proposes two new models that address these limitations. The models' cascading behavior turn out to match empirically observed behavior better than "SIR", a popular, existing model. (It is not made clear why other methods were not considered for the comparison.) The study is done on a 10% sample of the Twitter network over around 11 days. This is a great (and huge) data set, but it would have been nice with a discussion of whether the cascading behavior of a sample reflects the behavior of the full network.

Besides plotting the cascade size distribution, a Kolmogorov–Smirnov test is made to compare the model and true distribution. This makes sense, though I would have liked some more discussion of how to interpret Figure 6.2.

Finally, the thesis analyses the behavior of the SIR model and a variant (“SIR bar”) assuming that the social network is a directed acyclic Erdös–Rényi graph. Even though the results are "right", the cascade distribution follows a power-law, I am not sure how strong conclusions one can make about social networks, since they are known to look quite different from random graphs. It would have been interesting, but probably much more difficult, to analyze more realistic social network models, as well as to analyze the new dissemination models that are proposed in the thesis.

**Overall assessment**

Overall the thesis reflects research of high quality, using sound mathematical methods. I enjoyed reading it, and had already read some of the papers it is based on. The exposition is good but could be improved in some places (see detailed comments below). The only serious issue is that the thesis does not adequately discuss the state of the art, by not citing the paper of A. Wei. It may be that the author has simply not been aware of Wei's paper. I recommend that the relationship of Wei's techniques is discussed at the defense.

The thesis clearly fulfils the requirements for the PhD degree.

Rasmus Pagh

Professor

**Detailed comments**

- Why is the "fast query time" polynomial in d? Can you not always reduce to O(d/log n) problem instances in O(log n) dimensions, e.g. using the FJLT rotation (ref. [5]) together with your splitting aproach, in time O(d log d)?

- In section 1.2.1 it is not clear what metrics these results hold for.

- Section 1.2.3 says "If we were able to efficiently embed l1 into the Hamming space ... it would also give an algorithm for l2 and l1." Please discuss what is known about such embedings.

- References should be complete and unambiguous. For example "HT" in reference [12] is not a complete reference.

- The thesis should be spell checked and properly proof-read. A sample of typos seen:

Page 47: "optimal choice o w_i"

Page 48: "we prove results for special class"

Page 67: "can be approximated by power–law distribution"

Page 69: "might be to restricted", "interesting to introduce small number of"

Page 70: "Twiter, Flicktr", "folower", "A certain properties are"

Page 71: "radnom acyclic"

Page 73: "Infromation"