Nelly Litvak
Deptartment of Applied Mathematics
University of Twente
n.litvak@utwente.nl

Frank van der Meulen
Delft Institute of Applied Mathematics
Delft University of Technology
f.h.vandermeulen@tudelft.nl

Event STAR Outreach Day, 12 December 2014, Eurandom, Eindhoven

Networks & Big Data

Once a year, the NWO cluster Stochastics – Theoretical and Applied Research (STAR) organises a STAR Outreach Day, a one-day event around a theme that is of a broad interest to the stochastics community in the Netherlands. The last Outreach Day took place at Eurandom on 12 December 2014. The theme of the day was 'Networks & Big Data'. The topic is very timely. The *Vision document 2025* of the Platform Wiskunde Nederland (PWN) mentions big data as one of the six "major societal and scientific trends that influence the mathematical sciences". In 2014 a ten-year Gravitation programme 'NETWORKS' in mathematics was awarded by NWO. The STAR Outreach Day has presented an exciting opportunity to promote these topics and their numerous applications in life and science. Organisers Nelly Litvak and Frank van der Meulen look back on this event.

Networks are all around us, examples vary from neural networks in brains to Internet and social networks. Graphs provide a natural way to describe and study complex relationships within a network. As a first example, in social networks the vertices form the population that the social network models, while the edges indicate all the friendships in the social network. A second example is a graph in which nodes (vertices) represent diseases and symptoms. In this graph we then wish to infer causal relationships indicating which symptoms increase the probability of suffering from a certain disease. From a stochastic point of view, the problem amounts to estimating a dependence structure over a graph. The emergence of 'big data' often implies a huge number of nodes in the network, and computational feasibility of estimation procedures is achieved by imposing structure on the network.

Research on big data and networks is a new very active field of studies. The talks by six speakers at the STAR Outreach Day covered both applied and theoretically oriented research work. In this article we will briefly review these talks. Clearly, our summary is necessarily selective and may not do full justice to the presented research. However, we hope to give the reader an overview on the type of problems that were discussed during the day. Abstracts and further information is available at www.eurandom.tue.nl/STAR.

(i) 'Statistical Aspects of Tumour Heterogeneity' by Simon Tavaré. The first talk of the day was presented by professor Simon Tavaré, who is director of the Cancer Research UK Cambridge Institute and professor of Cancer Research (Bioinformatics) in the department of Oncology. His research focusses on cancer genomics. This field of research has seen a tremendous increase in acquisition of data which leads to many interesting new statistical problems.

In his talk professor Tavaré started with an explanation of the biological context for studies of cancer heterogeneity. From the associated statistical aspects that arise in its study we will elaborate a bit on one particular problem discussed by Tavaré (see also [2]). For this purpose we first explain a bit of terminology from systems biology. The complete inventory of all DNA that determines the identity of an organism is called its genome. The research in [2] deals with detection of mutations found in the genomes of tumour cells. Such mutations can be either 'somatic' or 'germline', where somatic indicates that the mutation cannot be transmitted to offspring. Discovering cancer-related somatic mutations is confounded by the presence of millions of germline mutations. Information consists of DNA sequence data from multiple spatially or temporally separated tumour samples from the same cancer patient. Interest lies in detecting somatic mutations from this information. Instead of performing multiple pairwise analyses of a single tumour sample and a matched normal, the authors consider all available samples jointly. Setting up a suitable statistical model is a nontrivial problem, and professor Tavaré explained how this can be done. In addition he indicated some computational difficulties in estimation of parameters for these types of problems.

(ii) 'The Structure of Critical Random Graphs' by Sanchayan Sen. The second talk was given by Sanchayan Sen. He has received his PhD from New York University, and has recently joined TU/e as a postdoc under the NWO Gravitation program 'NETWORKS'. The work of Dr. Sen is in the area of random graphs, which generates increasing interest in the Netherlands.

The topic of the talk was critical behaviour in random graphs. Consider the classical Erdős–Rényi (ER) model, where a graph has n vertices, and an edge between each pair of vertices is drawn with probability $p = \lambda/n$. A well-known phenomenon in such graphs is the *phase transition*: if $\lambda < 1$ then the graph will consist of small disconnected components; if $\lambda > 1$ then there will be a largest component of size $O(n)$. Both subcritical and supercritical regimes are well understood through the connection between random graphs and branching processes. Dr. Sen studies the *critical* regime, which is most challenging from a mathematical point of view. His approach is to view components of a random graph at criticality as metric spaces. Using this method, he has proved that a broad class of random graphs behaves similarly to the ER graphs.

Dr. Sen closed his talk with a list of open questions, among which the extension of his methods to 'scale-free' degree distributions with infinite third moments of degrees, that are often observed in real-life networks,

such as social networks or Internet. Although solving these problems will pose major challenges, it is obvious that the theory of random graphs has been developing very fast in the last years, and its power in explaining phenomena in real-life networks is ever increasing.

(iii) 'The Diffusion And Effectiveness of Cancer Awareness Campaigns on Twitter' by Tijs van den Broek. The last talk in the morning session was given by Tijs van den Broek, a researcher at TNO and a PhD student of University of Twente. His research focuses on the use of social media by activists to persuade firms to adopt sustainable policy and practices. Van den Broek and his colleagues at the University of Twente have recently received a prestigious Twitter data grant to analyse the diffusion and effectiveness of cancer early detection campaigns. With the data grant, the research team has received an unlimited access to Twitter's unfiltered archival data for a time period of two years. The research has just started, and there are more questions than answers. What drivers and barriers influence the diffusion process of Cancer Awareness Campaigns? And to what extent and under what conditions do online opinions about Cancer Awareness Campaigns lead to offline behaviour? Even filtering relevant tweets is highly non-trivial. For example, most tweets on 'cancer' are actually about the zodiac constellation in horoscopes.

The speaker has mentioned several problems that call for deeper mathematical and statistical analysis. For example, to what extent network structures that arise during an online campaign, predict offline behaviour? Extracting the knowledge from the large Twitter data set, with its underlying networks of followers and retweets, presents interesting opportunities for mathematical and statistical analysis. This is an example of an application that naturally calls for a collaboration between mathematicians and social scientists.

(iv) 'In-Core Computation of Geometric Centralities with HyperBall: A Hundred Billion Nodes and Beyond' by Sebastiano Vigna. The afternoon program opened with a talk by professor Sebastiano Vigna from the University of Milan. He is a renowned computer scientist who has worked on a large variety of topics, including algorithms for large graphs and theoretical/experimental analysis of spectral rankings such as PageRank.

Most of us have heard of the famous experiment by Stanley Milgram in the Sixties. Milgram found that distances between people in a social graph were surprisingly small. This phenomenon is often referred to as the 'small world phenomenon' or 'six degrees of separation'. In 2011 professor Vigna collaborated on the computation of the distance distribution of the whole Facebook graph with ≈ 721 million users and ≈ 69 billion friendship links. It has been the world's largest Milgram-type experiment! Imagine that Facebook users are vertices and the friendships are edges. Assume that we start from a randomly chosen vertex in this graph, and we want to reach another randomly chosen vertex by traversing edges along the shortest possible path. How many vertices shall we typically visit on the way? The result of Vigna and co-authors says that, on average, we will visit only 3.74 intermediate vertices. In other words, on Facebook there are just 3.74 degrees of separation.

In his talk professor Vigna has explained mathematical techniques behind these enormous computations, reported in [1]. It is common for mathematicians to study properties of graphs (including graph distances), when the number of nodes goes to infinity. However, when the graph is actually extremely large, all standard exploration-based approaches, that are used in theoretical analysis, are infeasible either in time, or in memory, or both. Professor Vigna and his co-authors developed a hyperball technique that employs probabilistic counters. Roughly speaking, each counter provides a randomized approximation for sizes of expanding hyperballs around each node. When many such counters are used, the distribution of the distance between the nodes can be obtained with sufficient accuracy. This talk was an impressive example of modern mathematical methods that enable efficient computations on an unimaginably large scale.

(v) 'Modern Day Causal Discovery: Challenges and New Applications' by Tom Claassen. The fifth talk of the day was presented by Tom Claassen, who is an assistant professor at the Institute for Computing and Information Sciences at the Radboud University Nijmegen. His research focusses on causal discovery. In his talk he explained how current state-of-the-art methods are in principle able to infer causal relationships from observational data. However, in the 'big data' era existing implementations are often ill-suited to handle the correspondingly large, high-dimensional data sets. Also, in real-world data many of the standard assumptions like multivariate Gaussian, or independent, identically distributed data do not apply, and standard causal algorithms may yield unreliable results or fail to run altogether. In the second part of his talk dr. Claassen discussed how existing methods can be adapted to provide more robust, informative, and realistic estimates on underlying causal mechanisms, and how to combine information from multiple data sets.

(vi) 'Big Networks and Data' by Ernst Wit. The final talk of the day was presented by Ernst Wit, a professor of Statistics and Probability at the University of Groningen. He discussed estimation of the structure of a sparse Gaussian graphical model. This is a popular mathematically tractable model for a big network, in which one expects only a small amount of nodes to be connected. Professor Wit discussed state-of-the-art methods for finding out which nodes are connected in the network. Instead of testing a large number of hypotheses (where each hypothesis asserts the presence or abundance of a connection), modern methods explicitly try to exploit sparsity. This has become a hot topic in statistics since the introduction of the 'Lasso' in regression analysis. Here, the basic idea is to estimate parameters by minimising a least squares objective function subject to a bound on the ℓ_1-norm of the parameters. The graphical Lasso is an implementation of this idea for Gaussian graphical models. In his talk professor Wit explained a novel, different, approach to this problem using approximate cross-validation and pointed out some of its advantages.

Acknowledgement
We would like to thank the speakers for contributing to this day. Some formulations are taken from their submitted abstracts.

References

1 L. Backstrom, P. Boldi, M. Rosa, J. Ugander and S. Vigna, Four degrees of separation, in *Proceedings of the 4th Annual ACM Web Science Conference*, ACM, 2012, pp. 33–42.

2 M. Josephidou, A.G. Lynch and S. Tavaré, MultiSNV: a probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples, *Nucleic Acids Research* (2015) 1–9.