# Algorithmic Questions for Data Science

Mary Cryan

31st October, 2016

# About myself

- faculty in the School of Informatics
- research on the Maths/Computer Science interface
- research topics include random structures, graph algorithms, randomized algorithms, and (long ago) learning of distributions.

# Models for Evolving Networks

*Define (random) rules for adding new nodes and edges to a network, and study the evolution of the network over time - we may be interested in analysing features like diameter, the degree distribution, connectivity, correlation of neighbourhoods.*

The goal is to design models whose long-term structure can be shown to be similar to real-life networks such as Web-like graphs, the Twitter network (topology or re-tweet structure), of the citation network (in academic publishing) etc.

# Experiments with Evolving Networks

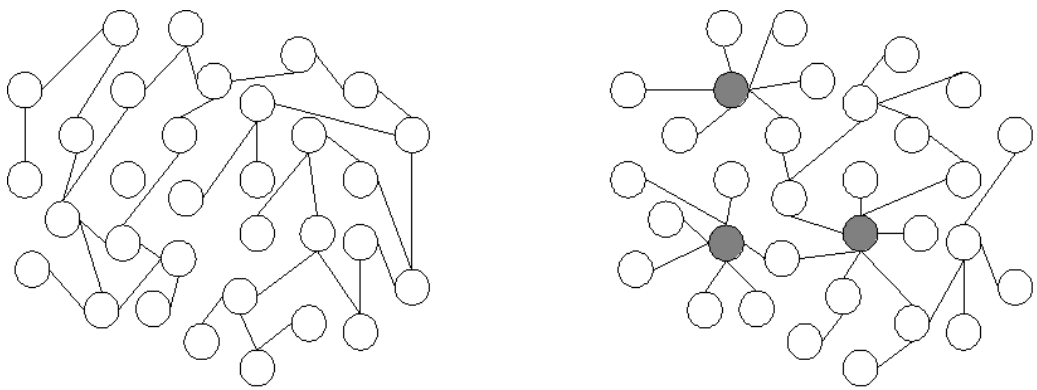
Early experimental work was done by Faloutsos, Faloutsos & Faloutsos "On Power-Law Relationships of the Internet Topology" (1999)

> *They analysed the physical topology of the Internet (on three dates in 1997, 1998), and observed that when the network nodes $v$ were grouped according to "common out-degree", and then the observed out-degrees $d_v$ were given a rank $r_v$ defined by the decreasing order of the* number of nodes *observed to have that out-degree, that the following "power-law" relationship held:*
>
> $$r_v \propto d_v^R,$$
>
> *for $R \sim -0.8$.*

Some of their claims/observations have been questioned.

# Random Models for Evolving Networks

A lot of theoretical work was carried out on the subject of designing random models for expanding networks which would generate networks with power law behaviour (scale-free networks) [Barabási & Albert, Cooper et al]. preferential attachment was a common feature of these models.



(a) Random network

(b) Scale-free network

# Evolving Networks and Data

Can mine large-scale data of a particular network to obtain statistics on many of the relevant structural properties of that network - and hence refine the design of a random model of evolution for that network. I had an MSc student do some experimental analysis of twitter data in the past, though it was a short-term project over a few months, and didn't include any theoretical analysis.

You may have some ideas of other data distributions that you would like to model. I'm happy to discuss.

# References

- "A brief history of Generative Models for Power Law and Lognormal distributions", by Michael Mitzenmacher, *Internet Mathematics*, 1(2), 226-251, 2004 (plus references therein): `http://www.eecs.harvard.edu/ michaelm/postscripts/im2004a.pdf`
- "On Power-Law Relationships of the Internet Topology", by Faloutsos, Faloutsos, Faloutsos; in *SIGCOMM 1999*: `http://www.cs.cmu.edu/ christos/PUBLICATIONS/sigcomm99.pdf`
- "Random Graphs", by Janson, Luczak and Ruciński, 2000 (maths textbook).
- Lots and lots of stuff from the "Stochastic Graph Models" workshop that took place at Brown Uni in March 2014.

# Differential Privacy

On the publication of large-scale statistical data, a crucial question asks how to "aggregate" data in such a way as to ensure that sensitive personal data of any individual, or group of individuals, cannot be computed (or approximated) from the data that is released. This question arises (for example) for the release of the "marginals" (row sums and column sums) of two-way contingency tables - while we might assume that no cells of the table can be inferred, performance guarantees need to be given before the values can be published.

# Data Privacy References

- "Privacy, accuracy, and consistency too: a holistic solution to contingency table release", Barak, Chaudhuri, Dwork, Kale, McSherry and Talwar, PODS 2007.
  `http://dl.acm.org/citation.cfm?id=1265569`
- "Differential Privacy: a survey of results", Cynthia Dwork, TAMC 2008.
  `http://research.microsoft.com/pubs/74339/dwork_tamc.pdf`
- Workshop on "Big Data and Differential Privacy" at Berkeley in December 2013.
  `http://simons.berkeley.edu/workshops/abstracts/78`

# To know more

- Happy to give reading-lists to students, or to supervise the MSc projects.
- I am teaching *"Randomness and Computation"* (level 11) this Spring. Anyone with interests in theory of Algorithms would enjoy this course.