Exponential random network models

organized by
Sourav Chatterjee, Persi Diaconis, Susan Holmes, and Martina Morris

Workshop Summary

1. Motivation

Large networks abound in the modern world. There has been an active development of social scientists who work with statistical models and innovative computer graphics. There has also been a healthy mathematical development around graph limit theory (Lovasz and many co-workers, ie AIM workshop, August 2011). The purpose of the meeting was to bring together these two groups and see if we could find common research areas.

One common focus were problems from the emerging area of exponential random graph models. If G_n is the set of all simple graphs on n vertices (undirected, no loops or multiple edges). A family of probability measures is defined via

$$P_\theta(G) = Z^{-1}(\theta)e^{\theta_1 T_1(G)+\cdots\theta_k T_k(G)}$$

Here T_i are graph statistics, e.g. $T_1(G)$ might be the number of edges, $T_2(G)$ the number of triangles and so on. The θ_i are real parameters and $Z(\theta)$ the normalizing constant. One standard task: given a fixed $G \in G_n$, choose θ to maximize $P_\theta(G)$, this is the maximum likelihood estimator: $\hat{\theta}$.

The social scientists had found that this usually useful tool doesn't work in many natural problems: there are θ values such that generating G from P_θ and then estimating $\hat{\theta}$ gives a very different value There are θ values such that small changes give rise to wildly different outcomes. These results have mostly been discovered empirically. Some theory is beginning to emerge [].One area was to explain/expand this theory to problems of practical interest.

A second focus:many of the theorists had really no real idea of what kind of data is out there. What kind of questions are of interest, what sort of specialty tools (e.g. R packages) are available.

The conference was a solid success. The groups found that they could talk. This was helped by a 'swing group' of statisticians who both prove theorems and analyze data. There were several hands-on tutorial sessions, both full and small groups, to explore the main tools– a suite of R-packages called `statnet`.

One theoretical breakthrough: one difficulty in computing the MLE $\hat{\theta}$ is the intractable normalization constant $Z(\theta)$. Chatterjee-Diaconis had a limiting approximation for this. It was unclear how large n needed to be for their approximation to be useful. The R packages allowed a first comparison to made with real numbers. For simple but important problems (e.g. the edge triangle model), the approximations seemed useful for n as small as 50. This comparison could only have been carried out at an AIM type workshop. The theory is tricky (strange topologies on infinite dimensional spaces, non-standard calculus of variations

problems). Similarly the programs had many interior functions that were very useful but not documented as standalone.

All participants were gratified and surprised by this first successes. This spurred us on to take the approximations seriously. This began at AIM and is being continued in several collaborations. It involves attempting to solve the infinite dimensional calculus of variations numerically. This seems to be working out; the solutions seem "nice" (piecewise constant functions with neat boundaries). It should be possible to prove some theorems.

Another success was the realization that several seemingly different models: Stochastic Block Models (SBMs), spectral approximation models, nodal models were the same. This means that theory and experience can be transferred. This realization took place over several days at the workshop. The two groups each had well developed models with theory, real examples and programs out in the world. No one had suspected the overlap (indeed essential identity) and once raised, both groups disputed it. Hoever, group interest and time for numerous discussions and clarifications in different vocabularies and notation won everyone over. The two group leaders were Ian Fellows and Stephane Robin agreed and their theories are starting to shuffle together.

A few other highlights of the meeting:

- Many of the models (exponential families) are based on statistical mechanics thinking where all the 'objects' (ie people in a study) are exchangeable. In reality, the objects have informative 'labels' or covariables that serve as annotation to the vertices, height, weight, age, disease history for instance. These must be taken into account in a serious data analysis. One of the physicists specialized in statistical mechanics present at the workshop told us that this was a crucial lesson. Seeing the many techniques that the applied statisticians have developed to incorporate covariates was a lesson to all. In principle, graph limit theory can incorporate covariates but this has not been actively developed. Enough people noticed that there is hope that this clearly posed research problem came to life.
- Perhaps the clearest open problem that came out of the conference is the need for a useful theory of sparse graphs. Graph limit theory, and exponential random graphs in particular work for dense graphs (the number of edges approximately of the same order as the square of the number of vertices). Real graphs have mostly bounded degree with perhaps a handful of high degree nodes. Since some of the exponential models are widely useful, this raises the question of the relevance of the limiting theory to applications.
- Another problem that emerged was the following: Many real graphs show three properties:
  (a)They have more triangles than would be predicted by simple Erdös Renyi random graphs.
  (b) Small words: the diameter (furthest away pair in a graph) is often quite small (e.g. six).
  (c) Power law behavior of degrees.
  The problem is this: there is no natural model for graphs that shows all three behaviors. There are models showing all subsets of two. This problem pointed out by David Aldous was discussed during the workshop and it may be that Sourav Chatterjee found a contender.

Many interesting case studies were presented including case studies on drawn from California schoolchildren interactions from the AD health studies. The detailed knowledge of some of the participants of the real data was very useful in trying to define the most important challenges in practical implementations.

For instance, a useful challenge that went from practice to theory was the following: the exponential graph model of the form

$$\log P_\beta(G) = \Psi(\beta) + \sum_{i=1}^{k} \beta_i T_i(G)$$

where Ψ is the log normalizing constant, the T_i are the graph statistics and the β are the parameters as before. The value of k is classically taken fixed. In quite a number of models (for instance the alternating stars model), the value of k grows with the number of vertices n, can this be accommodated?

Other progress at the workshop is hard to quantify, maybe making the co-authorship network before it occurred and after a few years would enable us to quantify its impact.