# MURI Update Meeting
# Multivariate Heavy Tail Phenomena: Modeling and Diagnostics
# October 16, 2015, Columbia University,

Sidney Resnick
School of Operations Research and Information Engineering
Rhodes Hall, Cornell University
Ithaca NY 14853 USA

http://people.orie.cornell.edu/∼sid
607 255 1210 sir1@cornell.edu

October 14, 2015

# Themes: Modeling, Analysis, Control & Design

A. Modeling

- Properties of network growth models with preferential attachment as models of social networks.
  - Preferential attachment yields node frequencies with in-degree $i$ and out-degree $j$ which have multivariate heavy tails in $(i, j)$.
  - Empirical degree frequencies are asymptotically multivariate normal for
    - undirected case (Resnick & Samorodnitsky)
    - directed case (Wang & Resnick).

    Elaboration of the MG central limit theorem.
  - Progress on understanding the structure of the extremes of node degree in undirected graphs. Method: embed in birth processes. Goal is to apply to inference of tail indices.
- numerical/computing/simulation/visualization software techniques for multivariate heavy tailed phenomena.
  - Simulation tools for generating data from such network growth models (Atwood, Roy).
    - Test algorithms.

    - ∗ Experiment with inference techniques.
  - Visualization in two and higher dimensions (Nolan) as an R-tool for using multivariate stable and max-stable distributions.
    - ∗ Multivariate meshes.
    - ∗ Numerical integration over simplicies.
    - ∗ Generalized spherical distributions–large parametric subfamily.
    - ∗ Models and software to allow computation with multivariate stable and extreme-stable distributions.
- Generation of multivariate heavy tails using linked systems of Poisson process driven stochastic de's (Gong, Towsley, Lu; Samorodnitsky).

B. Analysis

- inferential methodologies and statistical methods.
  - Data analysis: classical methods (glm, extreme value) seem effective but ignore discrete nature of the data. (Roy, Jiang, Davis, Zhang, Golnari, Wan)
    - With modest assumptions discrete mass functions with heavy tailed measures can be embedded in continuous densities. (Wang, Resnick)
    - Tail empirical process methods applicable to explain why node based data can be treated like iid (but with different limits).
    - Some progress overcoming censoring using MLE technique in one dimension. (Sun & Resnick)
  - Joint degree distribution estimation and random walks (Towsley)
    - Sampling algorithm estimates joint degree distributions of large complex graphs.
    - Uniformly place a population of random walkers and allow them to take a number of random walk steps in a coordinated manner.
    - Unbiased MLE estimator of tail estimates improves over uniform sampling of nodes.

- Dimension reduction:
  - ICA, PCA for large dimensions; testing on multivariate stable data (Nolan, Davis). Zhang applies PCA to cybersecurity separating malware traffic from benign background traffic in a time series of observed network traffic from a compromised host.
  - Use of method of projecting onto log-concave densities (Davis, Zou)
- Threshold selection in heavy tail inference
  - Hypothesis testing method reported by Samorodnitsky & Nguyen (1st year of project) for univariate and multivariate case.
  - Multivariate case: Distance correlation applied to find threshold where multivariate angle and radius are independent. (Davis & Wan).
  - Clauset method in one dimension of picking the threshold which minimizes the Kolmogorov-Smirnov distance between the fitted Pareto tail and the empirical tail. (Resnick & Sun, Wan)
- Models for highly dependent heavy tailed variables.
  - High dependence (eg. exxon vs chevron returns) can be measured by the smallness of the support of the

limit measure or the smallness of the support of the angular measure.

  - ∗ In cases where the support is small, a second heavy tail regime may exist which concentrates on the complement of the support and would increase accuracy of risk estimation.

- Reciprocity: Directed asymmetric networks usually have minimal reciprocity (percentage of node pairs linked by bi-directional edges).
  - In many real graphs or theoretical models the reciprocity $\approx 0$. Possible to construct a graph whose reciprocity is 1.
  - Have method for maximizing reciprocity for networks constrained to have prescribed in- and out-degree sequences.
  - Yields a characteristic allowing comparison of two graphs via *maximum achievable reciprocity*. (Towsley, Zhang, Jiang,...)
- Influence maximization in social networks:
  - Selecting the optimal seed set of influential nodes can be NP-hard.
  - However, the influence maximization problem can be solved with a scalable and provably near-optimal greedy algorithm. (Zhang, Towsley)

C. Control and design of complex systems with multivariate heavy tails: Application to cloud computing and mobile networks (Srikant, Shroff)

- Mobile application usage in smartphone wireless networks
  - Balancing fast response, battery usage, mobility patterns.
  - Source of heavy tails: pattern of app launch follows Zipf's law and run-times and inter-run-times are dependent heavy tailed.
  - Approach to control: Submodular minimization problem; approximation algorithm proposed and tested on trace data.
- Resource allocation problems; cloud computing.
  - Multiple dependent tasks with heavy tailed service requirements; parallel servers.
  - How to schedule data retrieval necessary to complete tasks and balance loads across servers.
  - Scheduling policies for several important classes of data downloading time distributions that are either delay-optimal or within a constant gap from the optimum delay performance.
  - Under a mean-field assumption, system performance is insensitive to the task-size distribution, beyond the mean

so the practical implication is that it is possible to design simple load-balancing algorithms for large systems that are robust against the impact of dependencies and heavy-tails in service-time distributions.

## Synthesis

- probability model (eg. pref attachment) $\longrightarrow$ simulation methods for model to collect data.
- new statistical methods & diagnostics $\longrightarrow$ evaluate on simulated data where correct answer is known because we know the generating mechanism
- Real data (eg. slashdot) $\longrightarrow$ apply statistical methods to get empirical properties. $\longrightarrow$ Compare such properties with simulated model data. $\longrightarrow$ Criticize model.
- Decent model allows study of control policies to achieve objectives.

## 2015 Octoberfest Schedule:

1. 9:00-9:05. Joe Myers: Government perspective & introduction.
2. 9:05-9:25 Resnick overview, introduction.
3. 9:25-9:55 Nolan
4. 9:55-10:25 Shroff
5. 10:25-10:55 Srikant

   10:55-11:15 break
6. 11:15-11:45 Towsley
7. 11:45-12:15 Zhang

   12:15-1:15 lunch [Catered by Columbia]
8. 1:15-2:15 Samorodnitsky
9. 2:15-2:45 Gong
10. 2:45-3:15 Resnick
11. 3:15-3:45 Davis
12. 3:30-3:40 Resnick Wrap-up: The coming year.

3:40:4:00 Government panelists caucus alone.

4:00:4:20 Government panel gives feedback to MURI team.