Multivariate Heavy Tailed Phenomena: Modeling, Diagnostics and Applications

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

http://people.orie.cornell.edu/sid
sir1@cornell.edu

MURI Natick Nov 21, 2016

November 15, 2016

Legs of the Stool

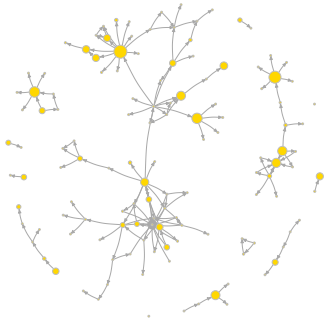
1. Mathematical modeling.
2. Data exploration, inference and network analysis.
3. Scheduling and control in multivariate heavy tailed environments.
4. Software.

1. Mathematical modeling

- Network growth under preferential attachment. (UMass, Columbia, Cornell)
- Generating mechanisms for multivariate heavy tails. (UMass, Cornell)
- Models for mobility, eg wireless networks. (OSU)
- Competition models under cumulative advantage; related to preferential attachment and generates games with heavy tailed durations. Cumulative advantage like preferential attachment may be linear or non-linear function of *fitness*. (UMass)
- Models to aid visual search. (UMass, Illinois)
  - Searching for suspects from a huge number of images.
  - Matching images when resolution is low in one image.
- Semi-parametric and asymptotic techniques for risk estimation. (Cornell)
- Methods for dimension reduction for multivariate heavy tailed data (ICA, PCA, ...). (Columbia, American)

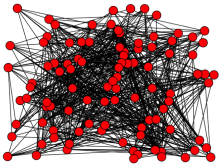
2. Data exploration, inference and network analysis.

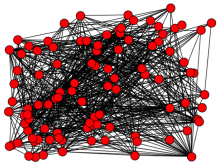
- Sampling issues: how to sample large networks under budget constraints to achieve estimation of summary quantities? (UMass, Columbia, Cornell, Minn)



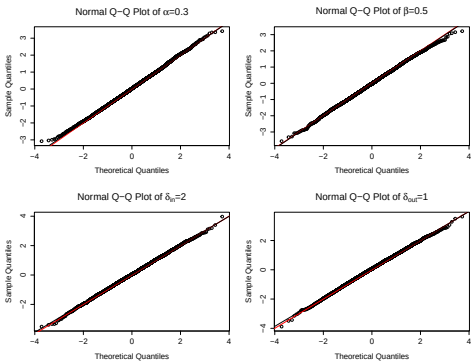
- Node classification: Questions are suggested by difficulties in sampling sensibly. Nodes may have multivariate characters: in- and out-degree, strengh of friendliness, strength of foe-y-ness, etc. (UMass, Cornell)

- Algorithms for matching large networks. (Illinois, UMass, Mn)
  - When are two networks similar?
  - Heavy tailed node degrees help the algorithm; match large degree nodes first.





- Calibration, simulation of a preferential attachment model. (UMass, Columbia, Cornell)
  - Efficiency loss from estimation using one snapshot vs network evolution;
  - stationary vs change-point analysis;
  - asymptotic semi-parametric methods (more robust) vs MLE parametric methods (more fragile but more accurate).

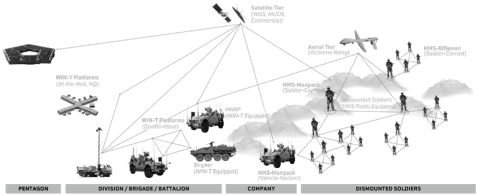


- Methods for exploring reciprocity (% of node pairs with bi-directional edges) in large directed graphs. How do reciprocal node pairs affect evolution and social structure? They form a core subgraph to which other nodes are attached. (Minn, UMass, Cornell)

- Mining geoMobility data (wireless users, bike-share, ambulance movement). Framework for study: EPIC. (Minn)
- Inference for heavy tail data (Columbia, Cornell)
  - Large values may be missing, lost, not-recorded.
  - Large values may be censored by rules. Must still do infence.
  - Limitations of minimum distance (Clausett) methods for threshold selection to decide what data has information about the tail.

3. Scheduling and control in multivariate heavy tailed environments.

- Emulating round robin scheduling in wireless networks. (Illinois)
  - Used for wireline where it is not sensitive to distributional properties of the file size distribution.
  - Military scenarios.
  - Goals to achieve: fairness, good throughput, small delay.



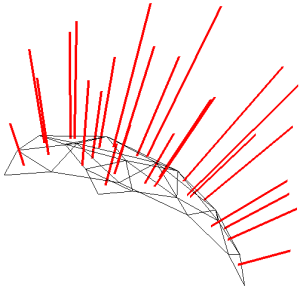
- Routing in cloud computing. (OSU, Illinois)
  - Massive data analysis is often needed to prepare for battle, to track adversaries, and for other military applications.
  - Routing algorithm that is insensitve to service time distribution.
  - Goals: Fast data retrieval, fast computation.
- Influence propagation in evolving networks. (Minn, OSU)

- Mobile data offloading: (OSU)
  - Cellular networks often highly constrained.
  - Suggests offloading some traffic to wifi or wired LANS.
- Context aware application scheduling for increasing battery lifetime and improving application response times in smartphones. (OSU)
  - Probability of launching an app follows Zipf's law.
  - The inter-run time and run time of an app follow multivariate heavy tail law.
  - Which apps held in memory and which should be closed?

4. Software.

Nolan (American) packages in R environment posted on CRAN (Comprehensive R Archive Network) at https://cran.r-project.org.

1. **mvmesh.** - a package to define and work with MultiVariate Meshes in n dimensions.
   - Define common shapes - hollow and solid spheres, simplices, rectangles, and tubes.
   - Plot these objects in 2 and 3 dimensions.
   - Define and manipulate shapes in higher dimensions.
   - Multivariate histogram functions
     - Count the number of data points in partitions of any of the above shapes,
     - Directional histograms that tally how many points are in a list of cones.

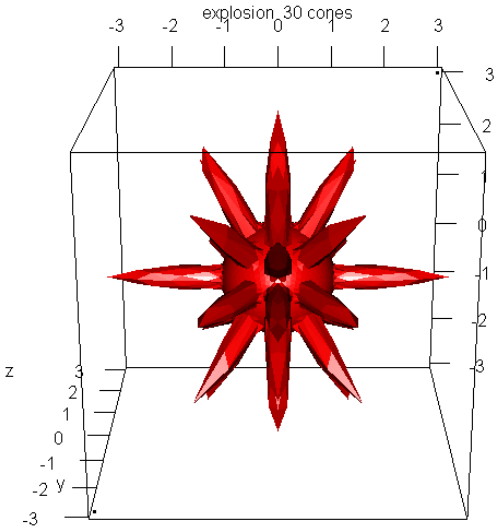


2. **SphericalCubature.** - a package to evaluate integrals over spheres in n dimensions. Gives exact formulas for polynomial integrands and adaptive methods for general functions.

3. **SimplicialCubature.** - a package to evaluate integrals over n dimensional simplices. Gives exact formulas for polynomial integrands and adaptive methods for general functions.

4. **gensphere.** - a package to define probability distributions that have level sets that are all scaled versions of a fixed contour.

   - Specify a large family of contours.
   - Define a probability distribution in terms of that contour and a radial decay function.
   - Arose out of previous Natick visit over attempts to model fragment dispersion from explosions.



4. **ecdfHT.** - a package for computing and plotting a transformed empirical cumulative distribution function (ecdf) for heavy tailed data.
    - Uses log-log transform with extremes pulled in to get visual diagnostic for heavy tails.
    - If plot suggests power law behavior, there are functions to estimate the exponent of the decay and the scale constant.
    - Multivariate generalizations are developed.
5. **mvevd.** (Under development:) Implements several dense classes of multivariate extreme value distributions.
    - discrete angular measure,
    - generalized logistic,
    - piecewise constant and linear angular measures.
    - Uses tools from the mvmesh and SimplicialCubature packages to work in dimension $n > 2$.

Contents