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 sidresnick@gmail.com

MURI Review: Cornell

September 30, 2013

1. Inter-related thrusts

Research effort organized into three inter-related and overlapping thrusts. Personnel overlap on the 3 thrusts.

1. Statistical methodology, computation, inference and diagnostics for heavy tailed data.
   - Assessing dependence. Model selection.
   - Estimating power law behavior from network (non iid) data.
   - Parametric subclasses of densities; computing limit measures; risk probability approximations; calibration of heavy tailed processes.
2. Mathematical modeling and implications.
   - Core theory and basic issues.
   - Model specific questions; development of methods applicable beyond the specific model.
   - Generating mechanisms for heavy tails.
3. System design, analysis and control in the presence of multivariate heavy tails.

- Impact of multivariate heavy tails on networks, computing systems, and human mobility; impact on performance metrics.
- Design principles and control strategies.
- Applications to
  - Coupled queues in cloud computing.
  - Coupled queues in wireless networks.
  - Deadline scheduling with outsourcing options.

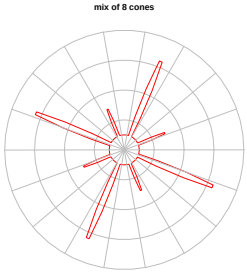
More detail.

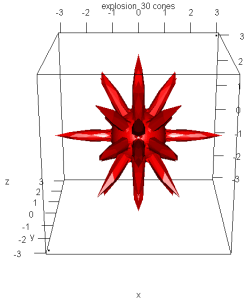
2. Thrust 1: Statistical methodology and diagnostics for heavy tailed data.

- Representations and characterizations of multivariate heavy tailed distributions (ongoing);

  – Parametric subclasses (Nolan) such as *generalized spherical distributions*; ongoing: increase flexibility.

  – Dealing with spectral measures of max or sum stable distributions in higher dimensions; discrete approximations or piecewise polynomial approximations; R-packages; ongoing: dimension severely restricted.





- Quantification of extremal dependence (Davis, Nolan, Resnick, Gena); risk contagion: do large values depend on each other; memory (see next Thrust).
  - Use of *extremogram* (Davis) to measure pairwise extremal dependence in (multivariate) heavy tailed time series. Measures the probability two lagged variables are simultaneously large.
    * Properties that overlap acf. Opens door to defining spectrum for extreme events. Asymptotic properties of extremal periodogram and its smoothed versions.
    * Extension to spatial extremes.
    * Uses for
      · heavy tailed time series model selection (eg. if the EDM$(h) = 0$ beyond some lag $h > q$, try MA(q));
      · Estimation of parameters in a space-time max stable process where observations are taken at irregularly spaced locations.
  - When extremogram indicates no extremal dependence, there may still be dependence on other scales (Resnick) that can be detected and used for risk estimates.
- Methods for estimating tail probabilities required in risk quantification; estimate $P[\boldsymbol{X} \in B]$ where $\boldsymbol{X}$ is a risk vector and B is a

remote region. Ongoing: high dimension.

- Data analysis of (in-degree, out-degree, ...) and neighborhood structure of social networks. Ongoing: Deal with censoring (Zhang, Davis, Resnick, J. Roy). Traditional graphic diagnostics help but why? (They were designed for iid.)
- Analysis of where the tail begins; threshold inference (Gena);
- Sampling distributions of diagnostic statistics; eg:
  - Extremogram.
  - Asymptotic normality for degree frequencies. Ongoing: Bo Jiang & RichardD. (Use martingale CLT?)
- Archiving, data collection, data access; currently informally shared: eg slashdot (Zhang), simulations from Krapivsky model (Atwood). Exchanged information about other archives such as SNAP.
- Software and Algorithms.
  - Display visual representations of heavy tailed data or level curves of heavy tailed densities (Nolan); identify preferred directions after intelligent scaling of multivariate data. Progress in 2 & 3 dimensions.

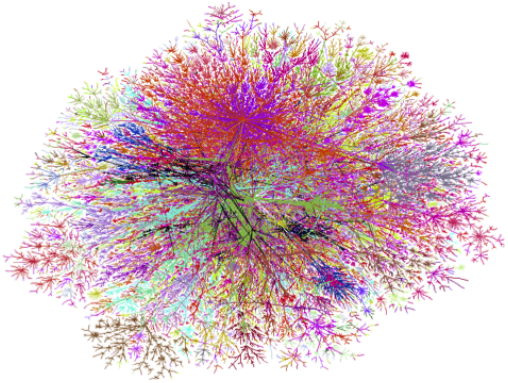
– Simulate data from heavy tailed densities where data concentrates in star shaped planar regions. (Nolan: ballistics lab in Natick).

– Compute quantiles of stable densities or densities of ratios of stable random variables (John Nolan). Ongoing: apply this to tests for independence in univariate heavy tailed time series (John & Sid)

– Simulate data from multivariate generalized spherical distributions (Nolan).

– Simulate heavy tailed processes:

  * growth models for social networks; eg. Krapivsky, competition.
  * Classify and learn directed networks;
  * How to generate? (James Atwood)

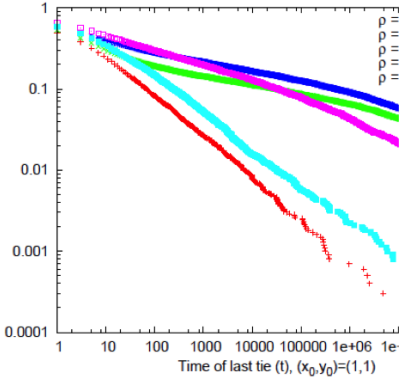


3. Thurst 2: Mathematical modeling and implications.

- Core issues: Problems arise in d-dimensions with $d > 2$ that do not arise in 1 dimension.
  - What does is mean for
    $$P[X = i, Y = j] =: f_{ij}$$
    to be multivariate regularly varying (have a power law)?
  - Embedding problem
  - Analogue of Karamata theorem for higher dimensions.
- Growth models of social networks: Krapivsky model, reciprocity model ({Don, Bo Jiang}∪{Zhi-Li} → {Richard, Sid}; Sid (back from Columbia) → Gena.
  - Gena: Explicit information on the joint generating function of the empirical frequency limits
    $$\{f_{i,j}, i \geq 0, j \geq 0\}$$
    for the relative frequency of nodes in the Krapivsky model with
    (in degree =i, out degree =j).
    Transform methods to determine the limit measure of regular variation.

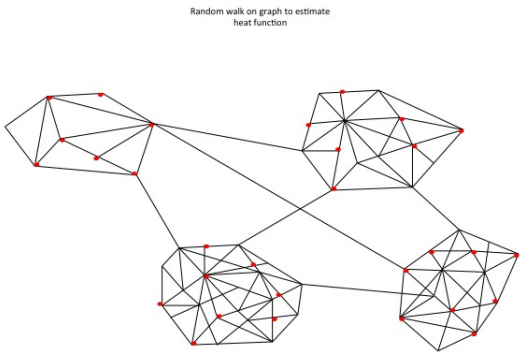
- Explicit computations provide examples of multivariate regularly varying mass functions which help when thinking about characterizations and solutions to the embedding problem.
  - For undirected graphs, vast literature in CS, Math'l prob relying on recursion, martingale methods, transform methods, embedding methods in birth rocesses or time changed non-homogeneous Poisson.
  - Less known for directed graphs where each node has multivariate characteristics; eg. (in, out, friend, foe).
- Simulated data aids in Thrust 1.

- Pairwise competition models based on cumulative advantage.
  - Account for *fitness* or current or cumulative wealth.
  - Investigate number of ties, and time of last tie.
  - Plot of tail distribution of time of last tie. $\longrightarrow$



- Tail data mining for high dimensional data.
  - In high dimensions, how do we discover where the tail or limit measure concentrates. Ongoing; some progress to estimate the support of the limit measure.
    - Where multiple heavy tail properties simultaneously exist on different scales and in different regions, there are multiple limit measures with different supports. These must be determined. For a particular risk estimation problem, an asymptotic regime must be determined. Basic understanding achieved but work ongoing (Resnick, Roy).
  - Is dimension reduction possible by an analogue of projection? (Gena, Nolan, Sid, Richard) PCA extensions: Richard has results on behavior of eigenvalues of random matrices with heavy tailed entries with dependencies allowed across rows and columns.

- Methods for determining when two networks are *similar*. (Gong, Zhi-Li)
  - Based on random walk over graphs that approximate diffusion on continuous medium;
  - experiments show good discriminative capabilities.
  - Ongoing: high dimensional data; which graph properties are continuous in the similarity measure. (Gong → Gena.)



- Asymptotic methods for estimating risk of remote regions and approximating ruin probabilities for risk and insurance. (Gena, Sid)
- Explicit influence of dependence & long memory on process tails and tail behavior of functionals such as the maximum of the heavy tailed time series or random field (Gena). Use of ergodic theory methods to describe long memory.

4. Thrust 3: System design, analysis and control in the presence of multivariate heavy tails.

- Coupled queues in cloud computing.
  - A model for decision making involving massive amounts of data.
    - A large job is divided into a number of parallel tasks (called Map tasks)
    - The results from the Map tasks are then aggregated to produce an output (called the Reduce task)
  - Multivariate heavy tailed vector:
    (number of Map tasks, duration of the Reduce task).
    Dependent components.
  - Goal: Understand the statistical properties of such systems and design good scheduling algorithms for such coupled queues (Map and Reduce).
  - Collaboration enabled by this MURI: Analysis and design of coupled queues in the large-system limit
  - Results:
    - Characterize dependence of the number of servers used in an infinite capacity system as a function of the multivariate Map/Reduce job size distribution

    - Low-latency scheduling algorithms for large capacity systems.
    - Detailed performance characterization in the heavy traffic limit.
    - A solution to the scheduling problem in which each task can request a specific amount of resources (such as CPU, memory, etc.) from the servers.

- Coupled queues in wireless networks.
  - Wireless resources are highly stressed.
  - Needs ways to off-load data from traditional cellular networks to other networks (e.g., WiFi zones, mmWave, etc.).
  - Observed times spent in and out of WiFi zones are dependent and distributed according to heavy tailed distributions.
  - For each user, the system can be modeled as a coupled WiFi and cellular queue, where the cellular queue serves traffic only after its time-out period has expired in the WiFi queue (reneging event).
  - Goal: Characterize the reneging probability waiting time in the WiFi queue.
  - Modeling: bivariate heavy-tailed on/off periods.
  - Results:

    - Optimal scheduling and outsourcing policies for single processor deadline scheduling.
    - Performance analysis under general arrival, jobs size, and distributions
    - Structure of optimal admission policy and heuristics.
    - Insights from numerical studies on the impact ofheavy tailed distributions on performance.
    - Optimal competitive ratio analysis.

- Deadline scheduling with outsourcing options.
  - Multiprocessor deadline scheduling.
    - Randomly arriving jobs with deadlines for completion.
    - Mechanisms to guarantee completion: local processing with outsource option and admission control
    - Applications:
      - Distributed situation awareness on the battlefield
      - Scheduling of cloud computing systems
      - Scheduling of large scale charging of electric vehicles
  - Goal: characterize statistical properties and performance; develop optimal/suboptimal online scheduling algorithms
  - Modeling: use heavy tailed distributions for job sizes, # of arrivals, and deadlines.

5. Visits and collaborations post Natick

- Fall 2012: Srikant visits OSU to confer with Shroff. This is followed up with visit by Srikant student Siva Theja Maguluri to Shroff in Spring 2013.
- January 8: Don → Sid at Columbia. Discussion of network growth models leading to further discussions with UMass group in February and with visit by Jiang to Columbia.
- Febrary 5: Sid & Richard → Don & Weibo & group at UMass.
- March 22: Sid → John Nolan at AmericanU.
- March 24-26: Lang Tong → Srikant in Illinois to collaborate.
- April 8: Bo Jiang (UMass) → Sid & Richard at Columbia.
- April: Nolan → Ananthram Swami (ARL) to discuss large scale networks and the sharing of data.
- April 26, 27, 2013: MURI team meeting at Columbia. Presentations and general discussions.
- Ongoing: May 2013. Conference call: John Nolan & Gena with Edan Ben-Ari & Phil Cunniff (Mech Eng) re model for explosive ballistics; software tools by Nolan and suggested model by Samorodnitsky.

- May 6: John Lavery → John Nolan at AmericanU.
- May 28, 2013: James Atwood (UMass) supplies Richard and Sid with a data generating tool to simulate the Krapivsky model and provide test bed for estimation methods. Result of discussions at April Columbia update meeting.
- June 5-7: Richard & Sid → Zhang at UMN. Slashdot data shared by Zhang.
- Shroff and Srikant jointly supervise OSU student Yousi Zheng on problems related to cloud computing. Weekly Skype conference calls every Thursday between Srikant, Shroff and Zheng. Result: Analysis and Design of coupled queues for cloud computing systems in the large systems limit.
- July 11-12 Tong → Ananthram Swami and Kevin Chan (ARL). Preceded by emails and phone calls.
- N. Shroff → Samorodnitski at Cornell in late summer 2013 to discuss large deviation techniques for heavy tail distributions. Goal: Formulate and investigate problems related to multivariate phenomena in social networks.
- July 24-26: Zhang → Towsley & Gong at UMass.

- August 26, 2013: Sid & Richard confer on estimation methods for network data at ISI Hong Kong.
- September 4: Conference call Bo Jiang & Towsley (UMass) with Davis (Columbia) and Resnick (Cornell); subsequent discussions and collaboration between Davis and Jiang on estimation.

Contents