MURI Update Meeting
Multivariate Heavy Tail Phenomena:
Modeling and Diagnostics
The Coming Year,

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 15, 2015
1. To do list

Inference

• Capitalize on the asymptotic normality results for node counts to do formal inference (estimation, model calibration).
  – Nodes with moderate degree may be modeled by glms (Roy)
  – Nodes with extreme degree modeled using heavy tailed methods.

• Study performance of sampling algorithms on data.
  – What is the sensitivity of the algorithm to
    * the number of random walkers?
    * the total graph size?
  – Need theoretical analysis relating performance to mixing times of coupled random walks on networks with heavy tailed degree distribution.
  – How sensitive is the algorithm to changes in heavy tailed dependence between in- and out-degree?

• Threshold selection.
  – Based on distance correlation.
∗ Test for multivariate regular variation and comparison of non-parametric and MLE estimation of the angular measure.
∗ Experiment with improving performance using weight functions tuned to heavy tails.
∗ Apply to time series and random fields.
∗ Try resampling methods for approximating the limit distribution.
  − Based on min K-S distance.
    ∗ How much error is made by assuming data is continuous when in fact it is discrete.
    ∗ Mathematical properties of procedure? Consistent? Asymptotically normal?
    ∗ Extend in higher dimensions? Does this require parameterization of the limit measure or angular measure?
  • Data reduction: Continue to develop theory of PCA & ICA for heavy-tailed data.
    − Establish theory for under-complete case when mean is infinite.
    − Derive algorithm for find the IC in the over-complete case.
    − Derive appropriate limit theory for over complete case.
– Apply some of these techniques to understanding hidden regular variation in high dimensions where it is necessary to estimate the support of limit measures.

• Data analysis on slashdot, Google+, . . . .
  – Evolution of Google+.
  – How do mutually connected components of Google+ evolve over a year period?

• Software:
  – Develop estimation methods for multivariate extreme value distributions and adapt the methods to a software package and release on CRAN.
  – Complete the PCA and ICA software; test on data sets.
Modeling

- Finish work on node degree as a function of graph size.
  - Embedding methods in birth processes.
  - Behavior of the tail empirical process based on degree of a node.
  - Application to
    * maximum degree node as a function of graph size.
    * inference using estimators such as Hill which are functionals of the tail empirical process.
  - Extensions to the directed graph case.

- Analyze mathematically reciprocity as function of graph size for standard models.

- Continue to develop generative models for multivariate heavy tail distributions aiming at understanding the source and implications of heavy tails in real data in complex networks and natural images.

- Graph exploration algorithm formulated as a multiarmed bandit with heavy tailed dependent rewards.

- Multiclass networks; ie, nodes with vector attributes where the vector may be multivariate regularly varying.
Design and Control

- Mobility problems:
  - How to learn the change of application usage pattern? Study as combinatorial multi-armed bandit.
  - How to exploit the similarity in application usage patterns of users?
  - How can we minimize staleness (i.e., the elapsed time since the last background activity) of applications that require content updates?

- Resource allocation.
  - Extend to more general distributions.
  - Understand the impact of dependencies among these distributions.
  - Load balancing aspects of resource allocation in queueing framework.
  - Develop a combined load balancing/scheduling scheme for data retrieval in cloud storage systems.
  - Continue to generalize results on insensitivity beyond the mean. Remove extraneous assumptions. Complete the work on the mean-field limit in full generality.