Maastricht University *Leading in Learning!*

Big Data in Finance

Alexander Grigoriev

School of Business and Economics
Sharing Success

Definitions

- Wiki: Big Data
  - Gartner's 3V-definition [2012]: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."
  - Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.
  - Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data.

My own definition

- ***Big Data*** as a dataset which cannot be physically **read** within any **reasonable time**.

- ***Reasonable time*** is the maximum amount of time that can be devoted to information processing when dealing with a specific problem.

Recent meetings

Big Data for Finance Summit

November 18 & 19, Boston, 2014

- Topics Covered Include:
  - Business Intelligence & Reporting
  - Planning, Budgeting & Forecasting using Data
  - Financial Analytics & Dashboards
  - Align Financial & Accounting Data
  - Becoming Strategic Advisors
  - Financial Planning & Analysis
  - Corporate Finance Technology
- Major directions:
  - Accuracy improvement
  - Speed improvement

We also do this and …

… Major directions:

  - Accuracy improvement
  - Speed improvement …

- Structural analysis of Big Data (this talk)
  - Key properties and their combinations
  - Data replication with simulations
  - Clustering
  - Ranking intangible objects

Bottom-up: Ranking intangibles

- Applicable to political parties, universities, journals, products, banks… Say, Alice considers a set of intangible investment projects…
- In a blog Alice counts the number of expert twits where a comparable pair of projects appears (**Big Data here**: sampling)
- Given a symmetric matrix of mutual appearances of project pairs, Alice has to find a ranking of projects that minimizes the total (or maximum) weight-distance over all pairs

Natural ranking defining assumptions

- Good investment projects are discussed more frequently than bad ones
- Good projects are compared to good projects, bad projects – to bad projects

Minimum linear arrangement

---	2	8	5
	---	3	6
		---	5

Project ranking	Weight-distance
1 2 3 4	2x1+8x2+5x3+3x1+6x2+5x1=53
3 4 1 2	5x1+8x2+3x3+5x1+6x2+2x1=49

Extensions

- Natural Language Processing:
  - “Project A is better than project B”
  - “Comparing A and B, the latter is preferable”

  This creates asymmetry in the matrix
- The matrix can be represented by a weighted complete graph -> If the graph satisfies some structural properties, e.g. bounded cut weight, can we solve the problem in polytime?
- Weighted LOG-distance is, actually, more realistic: towards the tail the distance gets less important
- Simulate the blogs and test the ranking using Monte-Carlo

Results on ranking

- Teodor Stoyanov (Business Intelligence MSc thesis)
  Chance prediction for French presidential elections of Sarkozy – Hollande: 49% to 51% resp.
  NLP based, corpus of 60 media articles, training set of 20 articles
- Andrey Kateshov (Economic & Financial Research MSc thesis)
  65 American universities + 60 liberal art colleges:
  the output is very close to US News ranking
  No NLP, blog = College Confidential
- Nick Mulder (Busines Intelligence MSc thesis)
  Top 10 car manufacturing brands
  To be defended...
  NLP based

Bottom-up 2: Big Data Clustering and Classification

Examples in Finance:

- Time-series clustering of stocks
- Fraud detection (more generally, Financial Credit-Risk Assessment)
- Clustering financial services (e.g. in London)
- Event clustering in insurance

Big Data definition reminder

- ***Big Data*** as a dataset which cannot be physically **read** within any **reasonable time**.

…

How can we cluster not reading the data???

Hint (from computational geometry):
to compute a geometric centroid with accuracy *$\varepsilon>0$* you need just a ***random*** sample of size *$f(\varepsilon)$*.

Results on clustering

- Omer Sami (Business Intelligence MSc thesis) Fraud Detection in an Istanbul Arts Auction: found a fraud classifier (function) for manipulative bids; two (may be three) historic manipulative bids are found

- Andrey Winokurow (PhD student) Generalized geometric 1-median problem with optimal price determination

Big Data simulations

- Consider social or bank or payment or project or any other network
- Assume Bob develops a “***method***” analyzing a large sophisticated project network satisfying some properties
- Bob wants to test his method on a large representative sample of networks
- Where to get this sample from?...

Simulated networks

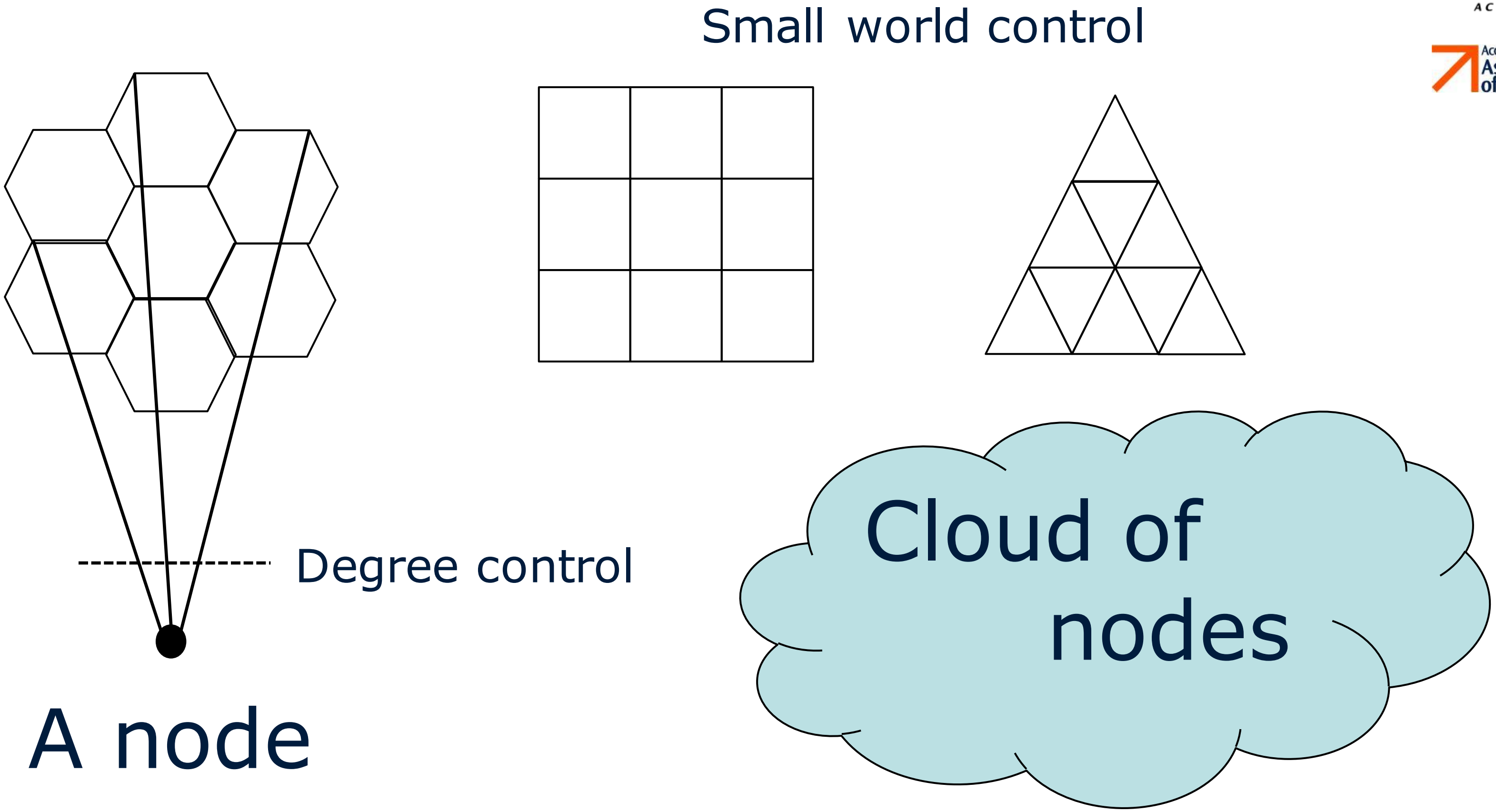
- Sizes: millions of nodes
- Properties:
  - Bounded degree (current analysis of Facebook & LinkedIn)
  - Clustering coefficients (current analysis of Facebook & LinkedIn)
  - Small world (my student's project, Agata Oleksy & Henning Reistenbach)

Simulated small world networks of bounded degree

- Let ***n*** be the number of nodes
- Let ***k*** be the average length of a shortest path (small world)
- Let ***l*** be the average degree of a node

Any idea how to create such a network?

Walls, grids and triangular tesselations



Small world control

Degree control

A node

Cloud of nodes

Simulating ANOVA output

- Given a data set on features of bank clients, Cyril wants to construct a linear regression for credit risks against all other features
- Cyril does it in Excel and obtains a classic ANOVA output (adjusted R^2, F value, intercept, coefficients, p-values etc)
- Now, for experiments Cyril wants to generate ***at random*** many new samples or one huge (Big Data) sample such that the output of ANOVA test will be similar to the one he has at hands

This project was given to one Master and to one PhD student & failed twice…
=> very challenging

Wrapping up structural and analytic problems in Big Data for Finance

- Ranking
- Clustering & classification
- Data replication & simulations for networks
- Data replication & simulations for regressions

Thanks!