Entropy-based Selection of Graph Cuboids

Dritan Bleco
dritanbleco@aueb.gr

Yannis Kotidis
kotidis@aueb.gr

Department of Informatics
Athens University Of Economics and Business

Grades 2017 - Chicago

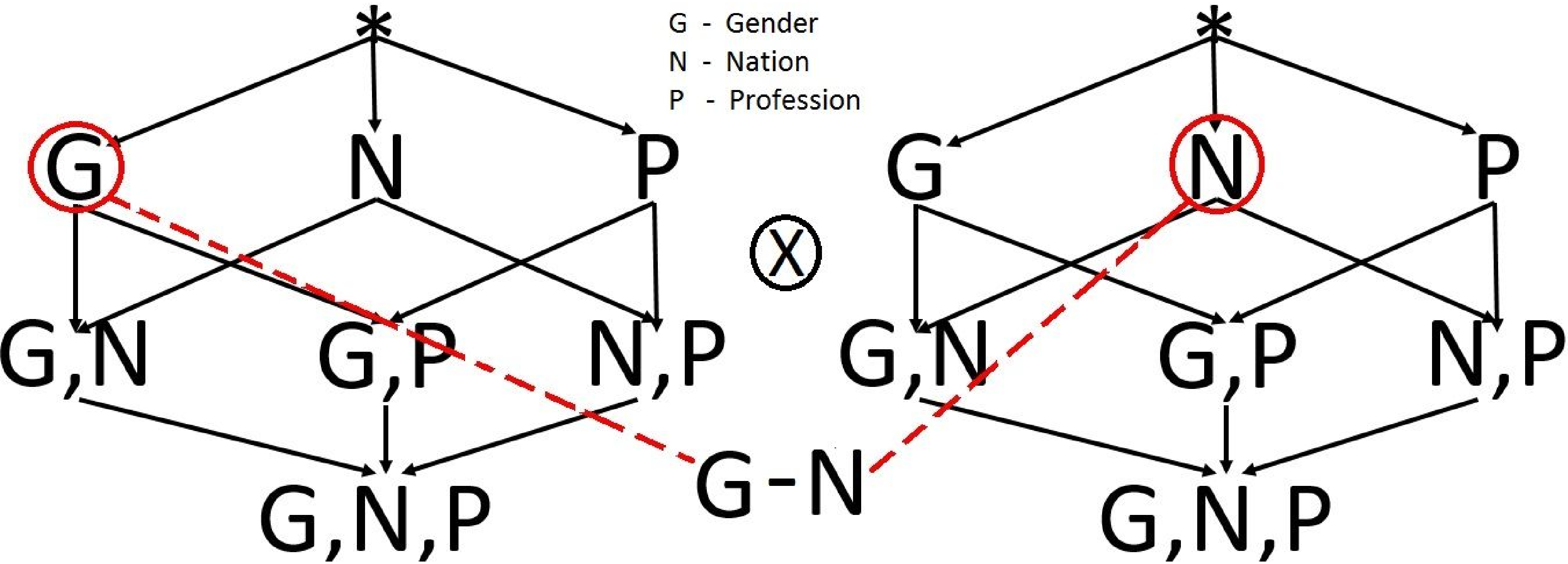
Outline

- Motivation
- Graph Cube
- Entropy – main concepts
- External and Internal Entropy
- Experiments
- Conclusions

Motivation

- Recent interest on big graphs with attributes at node/edge level
  - Running example: social network with 3 attributes on nodes: Gender, Nationality, Profession
- Graph cubes enable exploration of graph datasets by considering all possible aggregations among the node/edge attributes
- Our techniques aim at selecting subsets (called cuboids) from very large Graph cube by utilizing information entropy

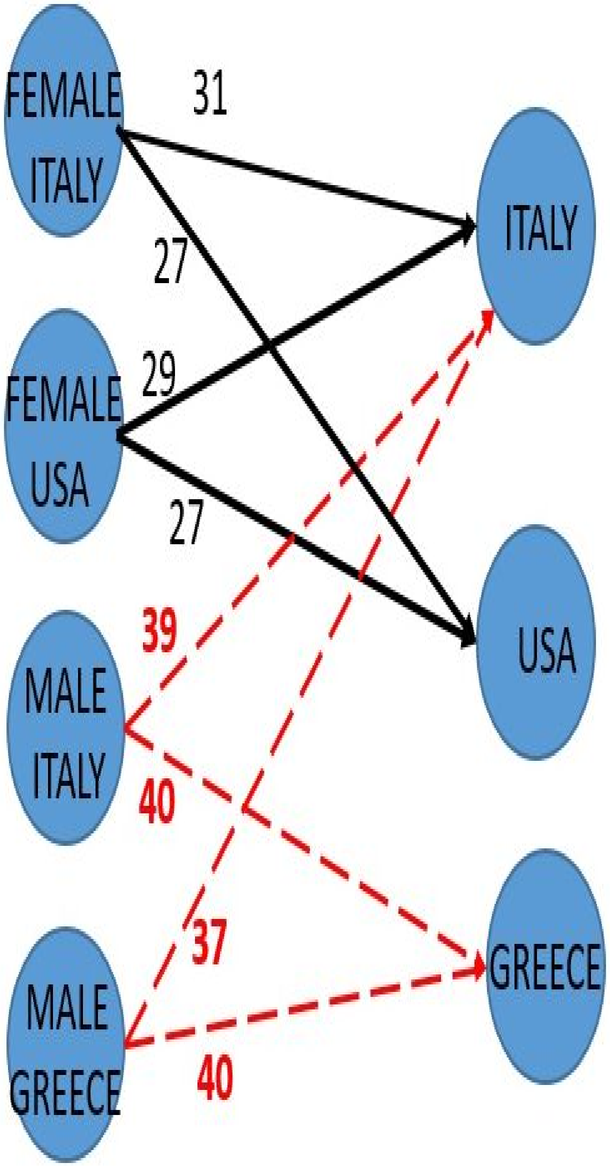
The Graph Cube



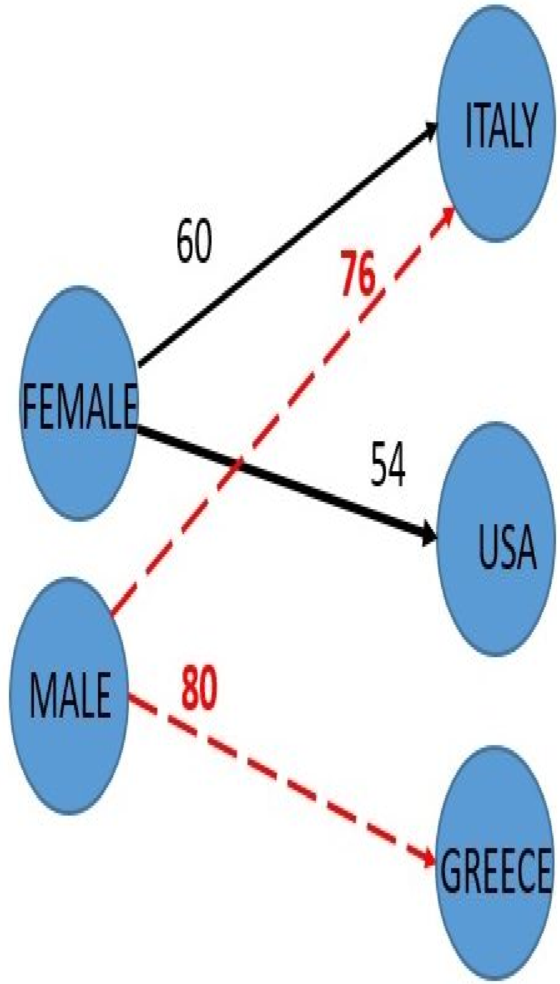
The Graph Cube : Cartesian Product of two cubes Starting (2^n) and Ending (2^n) Data Cube (2^{2n} cuboids in total)

Dimensions : Grouping attributes used in the analysis

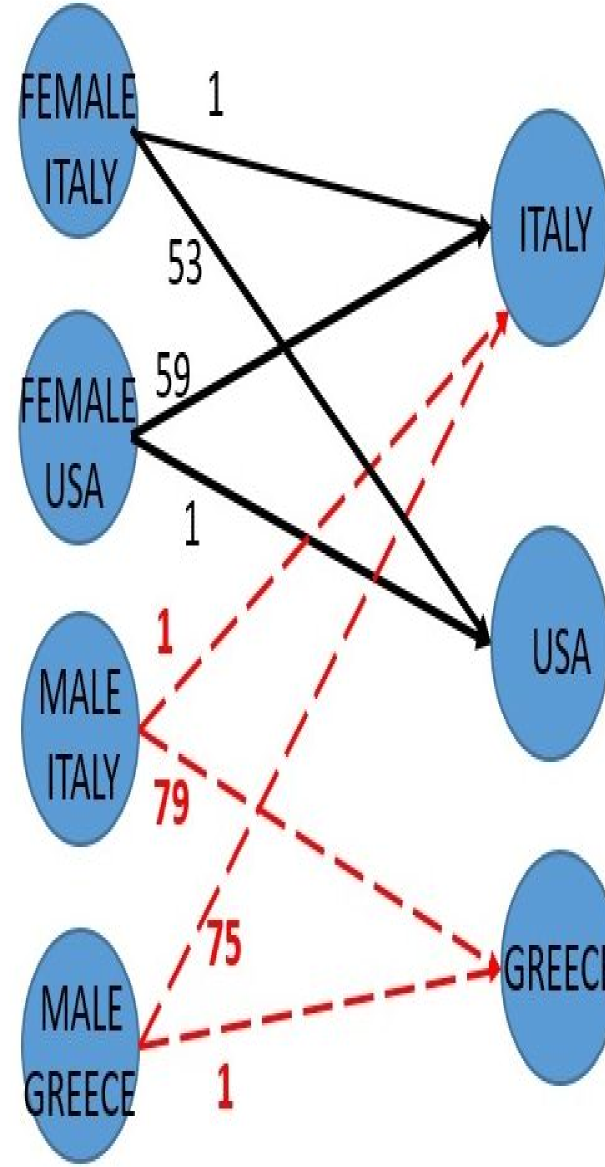
Cuboid : The result set of a particular grouping on the selected dimensions



Gender, Nation - Nation



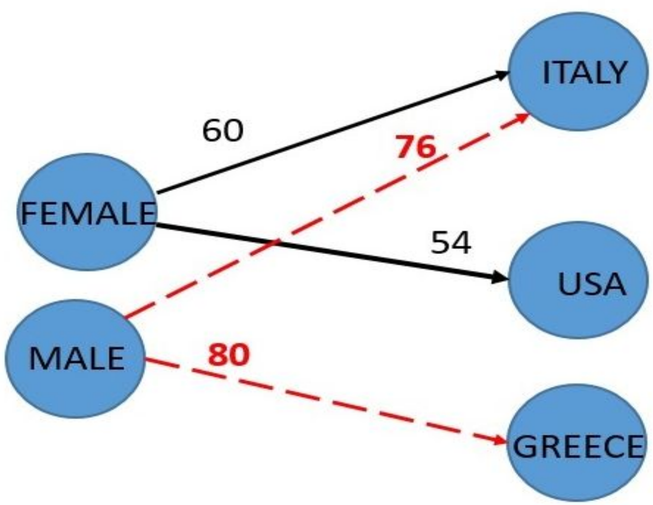
Gender - Nation



Gender, Nation - Nation

Cuboid Dual Representation

- Cuboids in graph cube may be represented as relations
- Relation schema contains attributes of starting and ending nodes and the computed aggregate



Gender - Nation

Record		Cardinality
$gender_s$	$nation_e$	
male	Greece	80
male	Italy	76
female	Italy	60
female	USA	54

Entropy - Navigating Graph Cube

- Analysts attracted by skewed data hidden in peaks and valleys
- Information Entropy or Shanon Entropy captures the amount of uncertainty

$$p(a) * \log p(a)$$

  – Increases when data are uniform
  – Decreases when there are high peaks or irregularities
- We distinguish External and Internal Entropy

External Entropy

- Cuboid C_i with m number of records in dual relation DC_i

$$eH\ (C_i) = -\sum_{j=1}^{m} p(a_j) * \log_2 p(a_j)$$

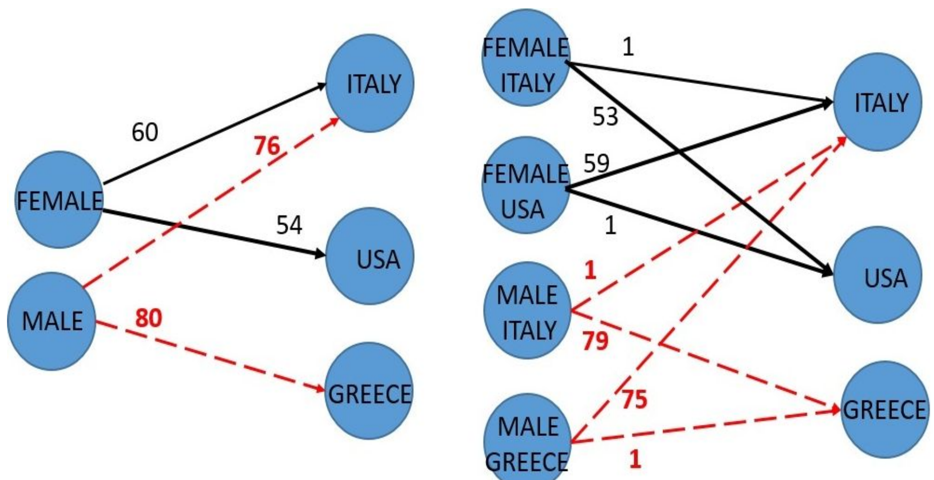
$gender_s$	$nation_e$	a	p(a)
male	Greece	80	80/270
male	Italy	76	76/270
female	Italy	60	60/270
female	USA	54	54/270

- Drilling down from Cuboid C_i - *parent* to Cuboid C_k - *child* adding attribute A with d_{max} distinct values
- External Entropy Rate

$$eH_{rate}(C_k, C_i) = \frac{eH(C_k) - eH(C_i)}{eH^i_{max}(C_k) - eH(C_i)}$$

Drill down (C_i, C_k) omitted if

$eH_{rate}(C_k, C_i)\ >\ eH_r$ (threshold)

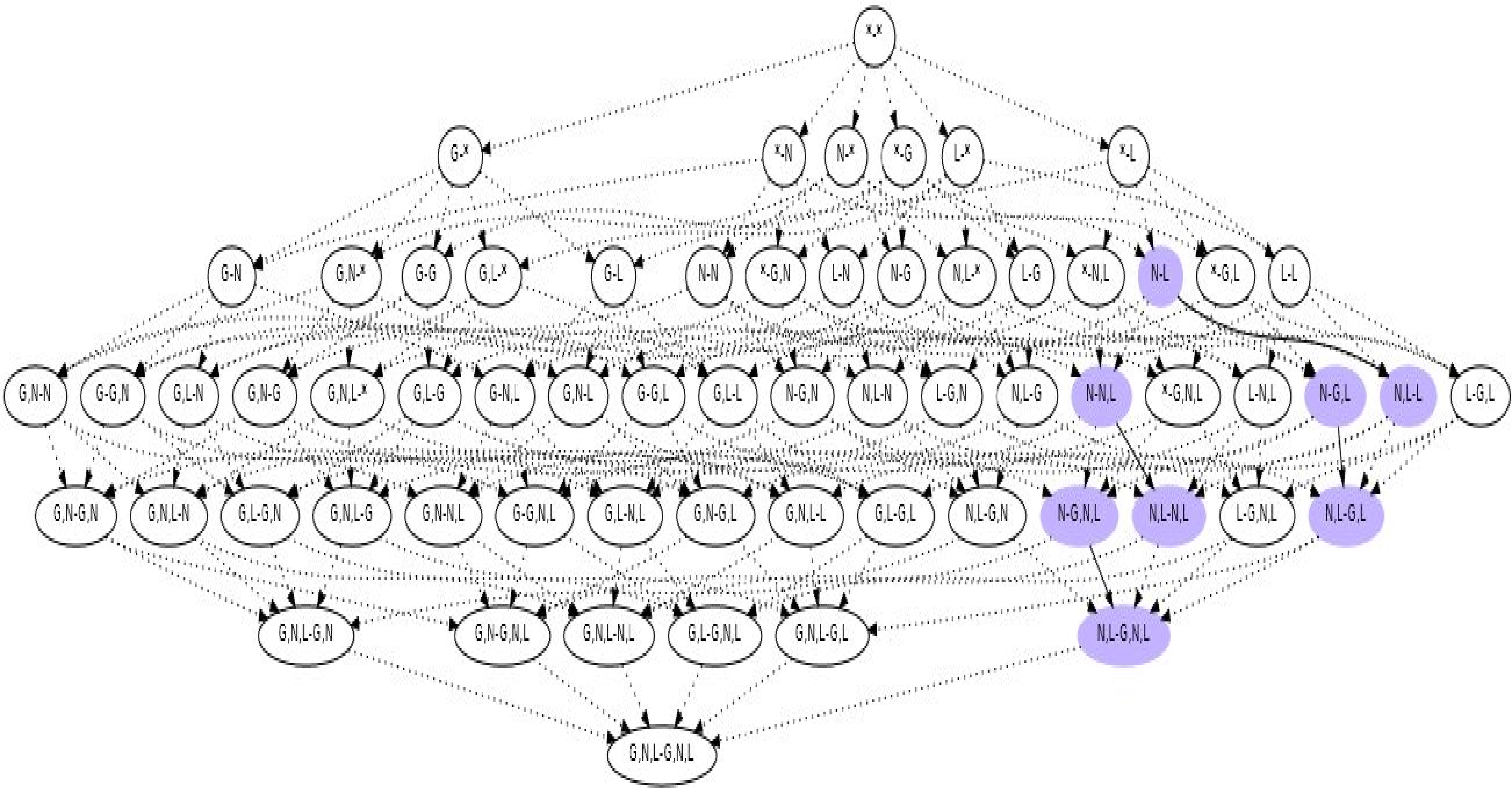


Gender - Nation

Gender, Nation - Nation

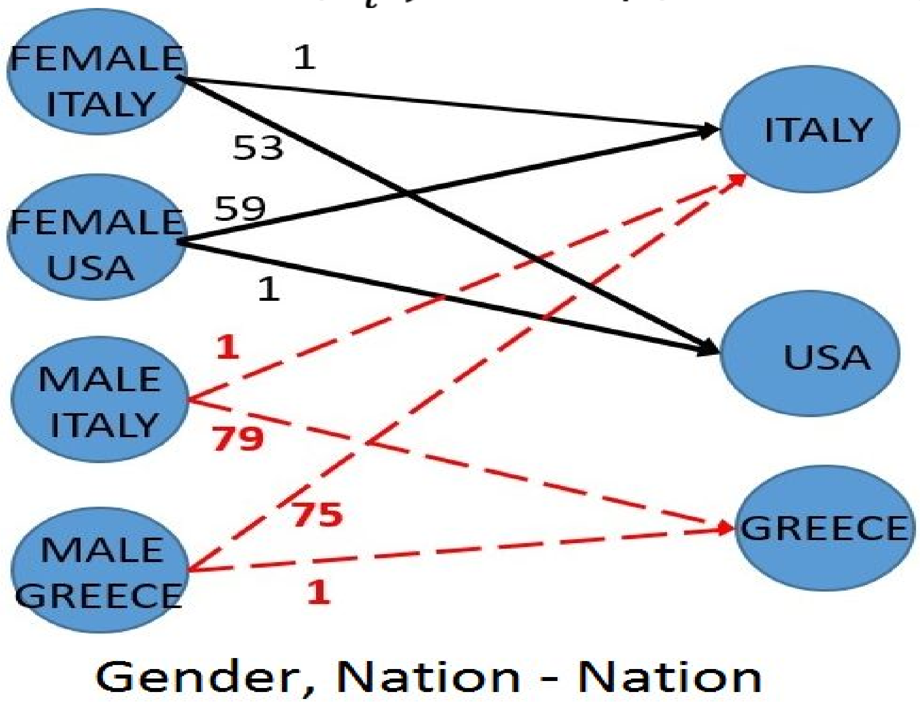
External Entropy

- Pruning Drill downs using External Entropy Rate



Internal Entropy

- Starting/Ending Internal Entropy Rate
- $sIH_{rate}(C_i^y) = \frac{sIH(C_i^y)}{sIH_{max}(C_i^y)}$
- Select prominent trends within cuboid
  - $sIH_{rate}(C_i^y) < sIeH_r$(threshold)



Experiments

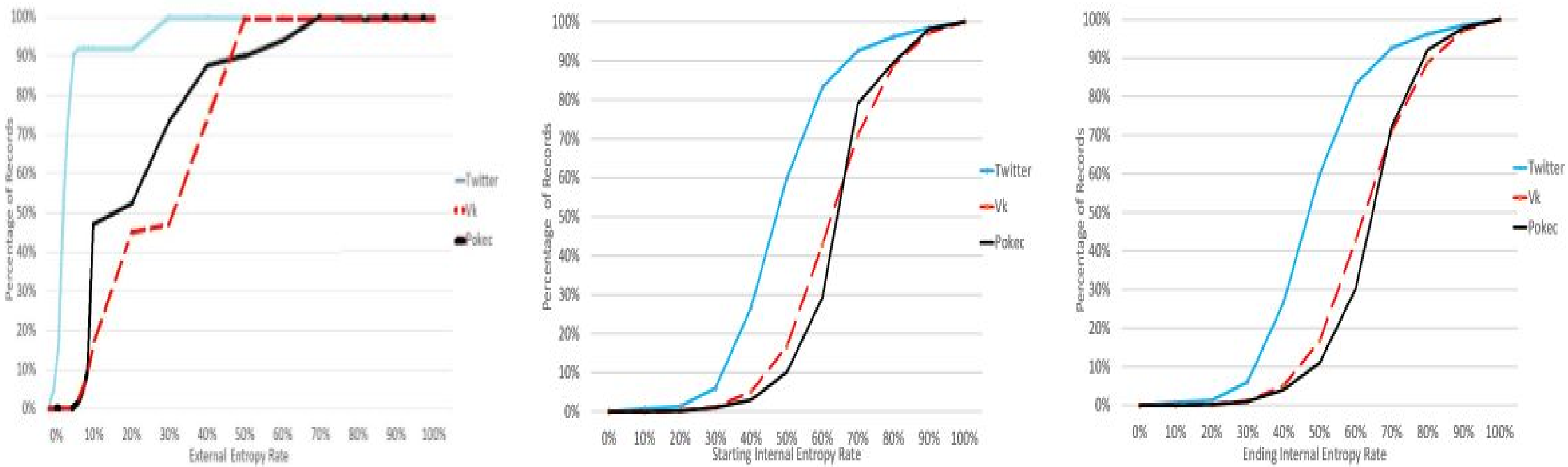
- Graph records from three real datasets
  1. Twitter: Crawled by our team
  2. VKontakte : The largest European on-line social network service
  3. Pokec : The most popular on-line social network in Slovakia

	Twitter	VK	Pokec
Profiles (nodes)	34M	3,9M	1,6M
Relations (edges)	910M	493M	31M
Number of Attributes	3	5	6
Number of Cuboids	64	1024	4096
Graph Cube Records	4M	362M	66,3B
Graph Cube Size	143MB	235GB	1.58TB

- Experimental evaluation using a Cluster
  - with 4 desktop each 4GB ram and 2T HDD
  - Intel i7-3770 3.40 GHz8
  - 8 VMs – one master and 7 slaves
  - Implementation using Apache Spark

Experiments (2)

- External and Internal Entropy Statistics

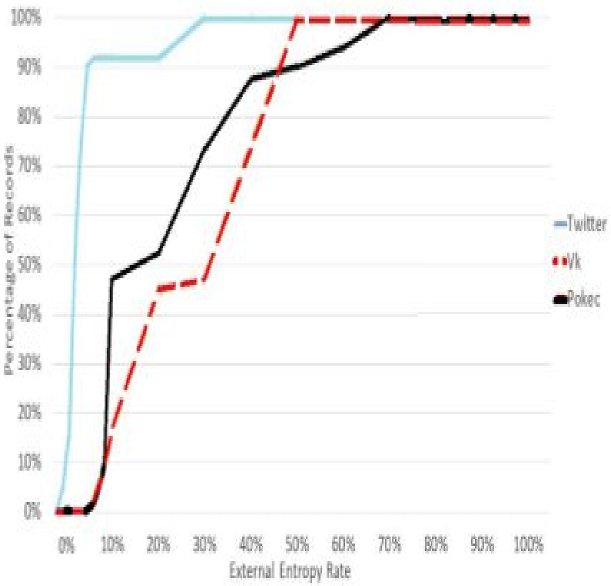


(a) Scaling external entropy rate (b) Scaling starting internal entropy rate (c) Scaling ending internal entropy rate

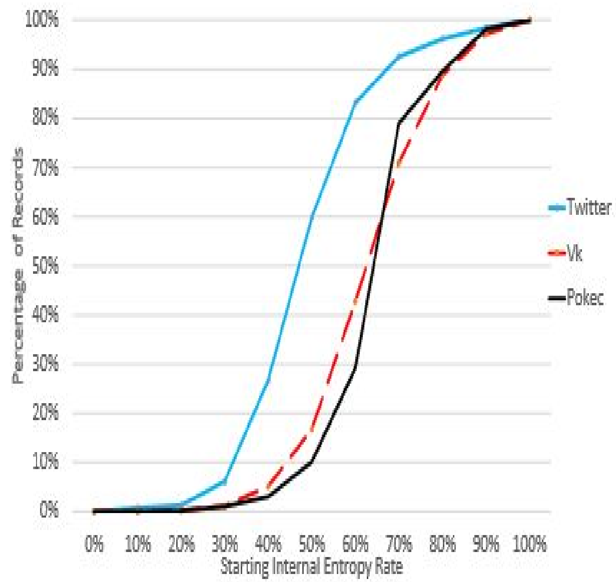
- Twitter : eH_r = 3.5% - 14% of dataset remains
- VK : eH_r = 10% - 17% >> >> >>
- Pokec : eH_r = 9% - 13% >> >> >>

Experiments (3)

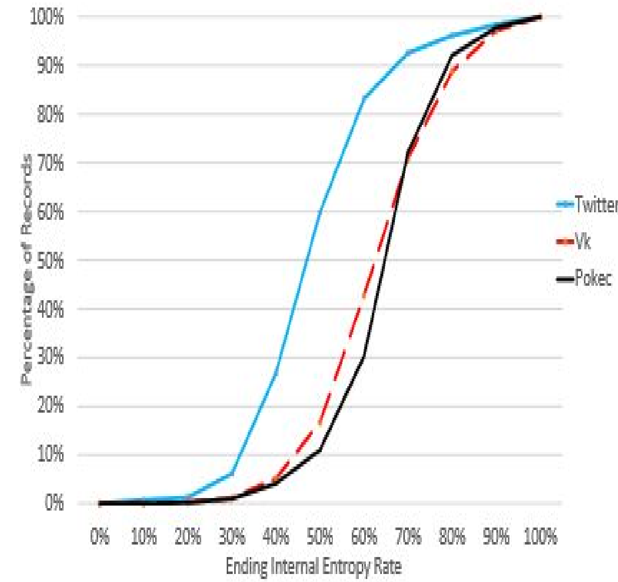
- External and Internal Entropy Statistics



(a) Scaling external entropy rate



(b) Scaling starting internal entropy rate

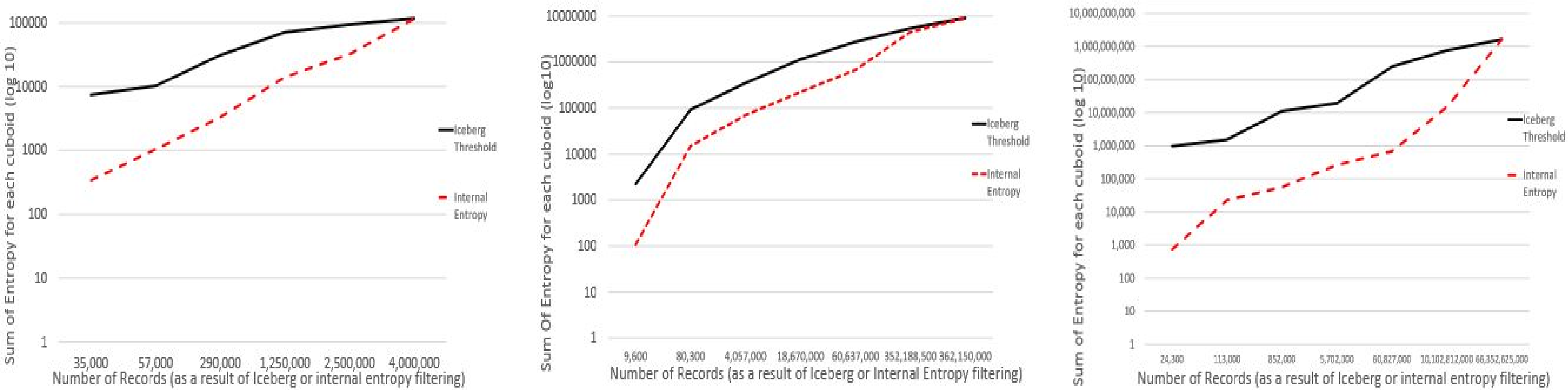


(c) Scaling ending internal entropy rate

- Twitter : siH_r = 10% - 0.70000% of dataset remains
- VK : siH_r = 10% - 0.00300% >> >> >>
- Pokec : siH_r = 10% - 0.00200% >> >> >>

Experiments (4)

- Iceberg graph cube vs Entropy



(a) Twitter dataset (b) VK dataset (c) Pokec dataset

- Compute the Iceberg graph cube for different minimum support and adjust Internal Entropy retaining the same number of records

- Compare the resulting subsets of the graph cube in terms of the sum of entropy retained in them.

Conclusions

- We presented a framework of graph cubes representing them as Cartesian product of independent data cubes on the starting and ending nodes of the graph
- Addressed the enormous size and complexity of the resulting graph cubes by proposing an analysis process that steers users towards interesting parts of the resulting aggregations.
- Our methods utilize intuitive entropy measures that help locate skewed associations
- Experimental results validate the effectiveness of our techniques and indicate that real graph cubes do contain interesting trends
- Our proposed optimizations enable us to manage graph cubes containing billions of records

Thank you,

Questions?