A retrospective on 20 year of Artificial Intelligence in Internet measurement and cybersecurity research

Kavé Salamatian, University of Savoie-Mont Blanc

IA revolution

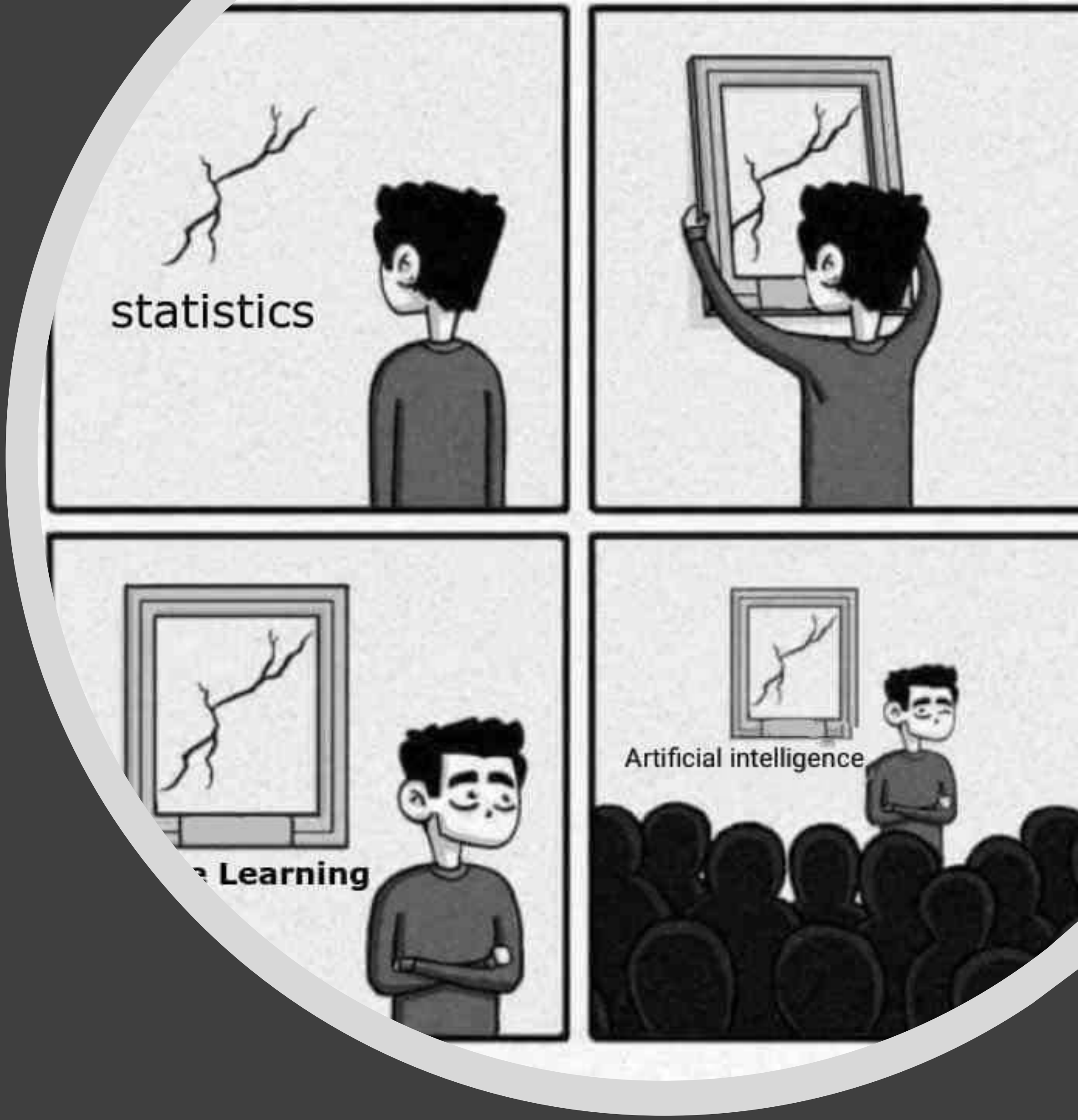
- **Robotsky**: You have
- **Robotsky**, **Sparx**, **M** tionary Front!
- **Sparx**: I'm Sparx, th
- **Robotsky**: And I'm R
- **Mike the Fridge**: An dge.

Some history …

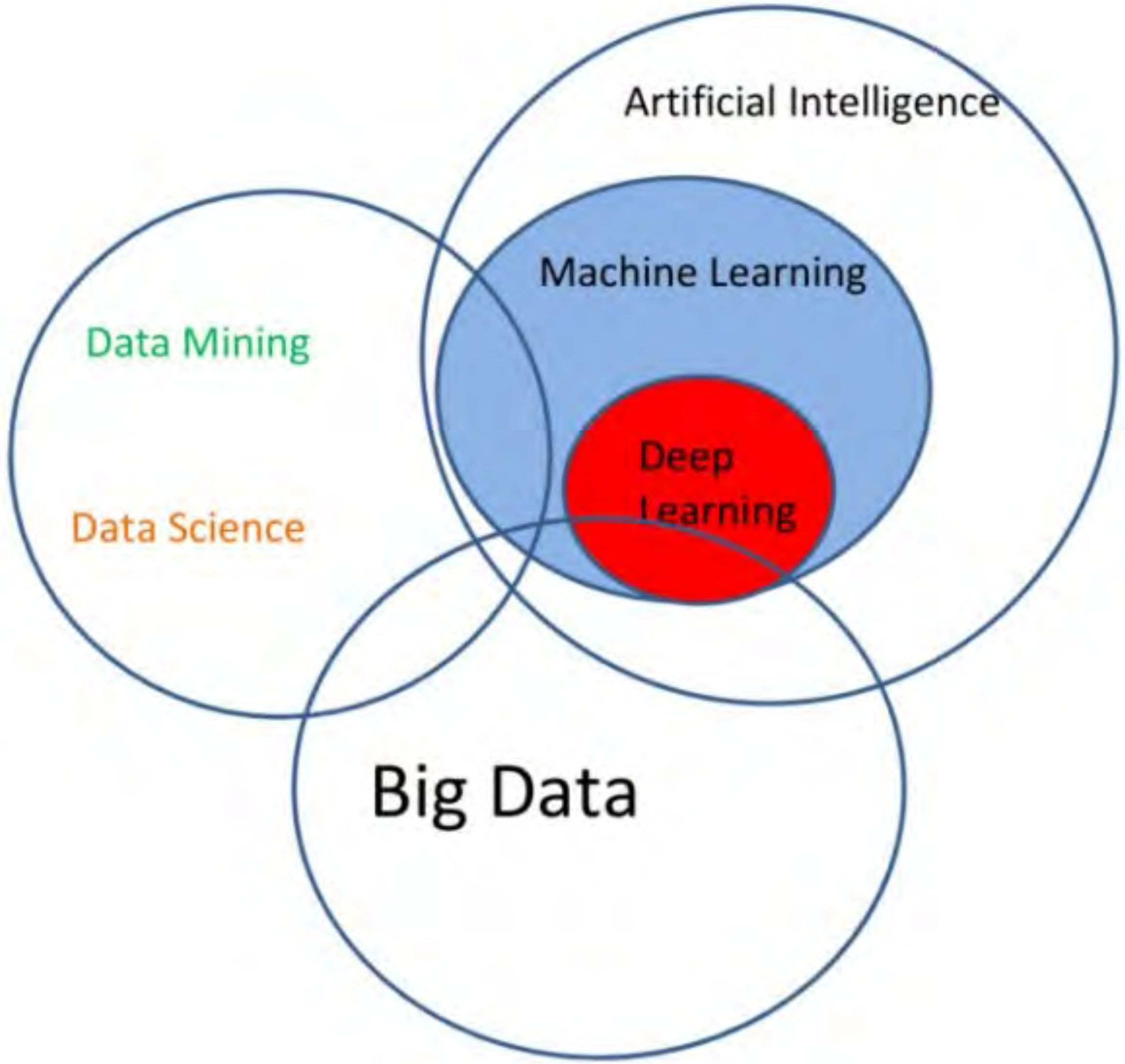
- **1969** — Arpanet
- **1982** — First virus : Elk cloner
- **1983** — TCP/IP
- **1988** — First worm: Morris worm
- **1988** — First Firewall
  - First stateful firewall, 1990
- **1997** — First DoS attack : 1997, during DEF CON
  - First DDos : 1999, Trinoo
- **1997** — First large scale Internet Measurements: 1997, Vern Paxson
- **1998** — Snort intrusion detection System
- **2002** — In 2002 first Internet measurement Workshop
  - Out of 39 papers, 12 about network anomalies …

What is AI ?

- Wikipedia : Artificial intelligence (AI) is intelligence demonstrated by machines.
  - the term "artificial intelligence" is applied when a machine mimics "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving »
- *“When you’re fundraising, it’s AI. When you’re hiring, it’s ML. When you’re implementing, it’s logistic regression.”*



AI, machine learning, data mining



Any fool consider himself as intelligent

Danish folklore

What is cybersecurity ?

- Communication security
  - Encryption, authentification
- Physical security
  - Firewalls, resilience
- System security
  - Malware, virus
- Software security
- Network security
  - Routing, in network detection
- User security
  - Social engineering
- Attacks intentions
  - Geopolitics, cybercrime
- Policies, regulations

"THEY WERE WAY AHEAD OF US IN PASSWORDS."

Machine learning in cybersecurity ?

- Model based
  - Traffic generative models
    - Queuing theory, Poisson models, Erlang, Markovian, self-similarity
  - Behavior models
    - State machine, Markovian, Latent state
- Inverse inference
  - Having an empirical traffic what are the parameters of the model
    - Moment methods, Maximum Likelihood, EM methods
- Model based anomaly detection
  - Calibrate a model of normal behavior
  - Detect divergence from normal behavior
  - Raise an alarm when divergence large
- Extension to non-parametric models

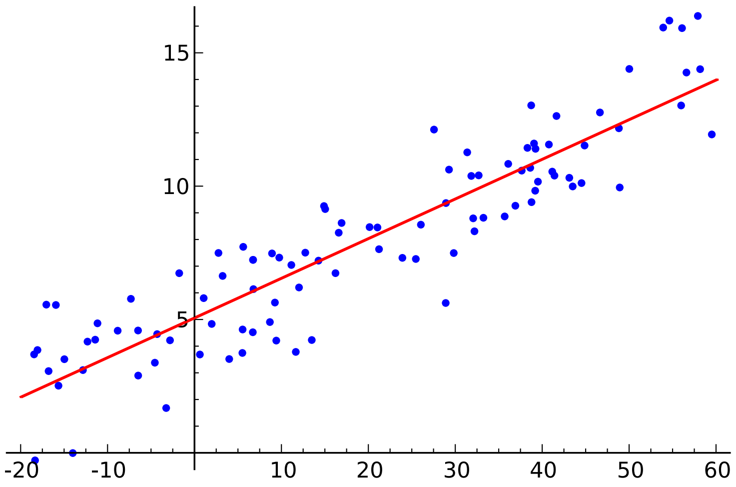
Data mining approaches in cybersecurity

- Association rules
  - For Anomaly extraction
  - For Rules extraction
- Sequence prediction
  - Similar to machine learning but with a sequence model
- Fingerprint extraction
  - Virus/Worm detection
- Log analysis
  - Natural langage processing, LDA,

Big data and cybersecurity

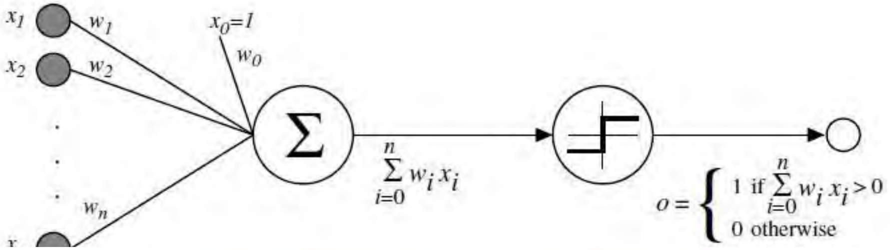
- One day of packet headers= 12 Tbytes of data
- 2 days of DNS data in China = 20 Tbytes of data, 72 Billions records
- AS level graph analysis : one 68 k nodes graph per mins over 50 days
- Social network graph: 300 k nodes graphs
- Fraud detection : 75 k nodes graphs

A critical history of deep learning

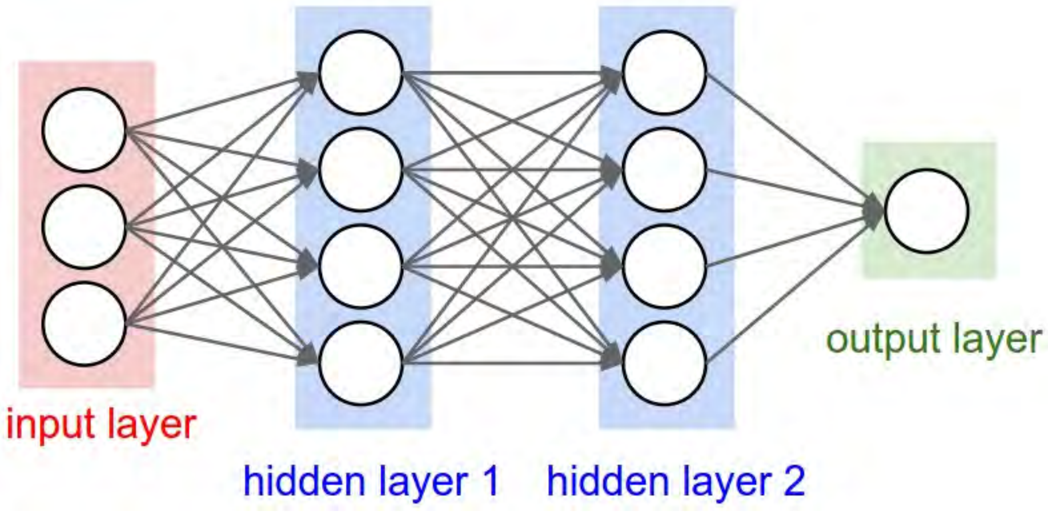


- The Centuries Old Machine Learning Algorithm

- The perceptron: 1957

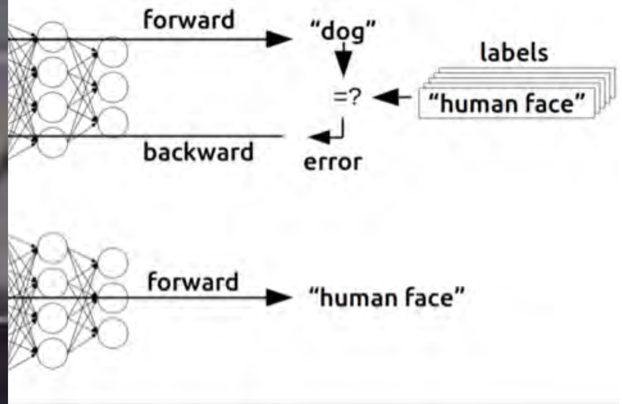


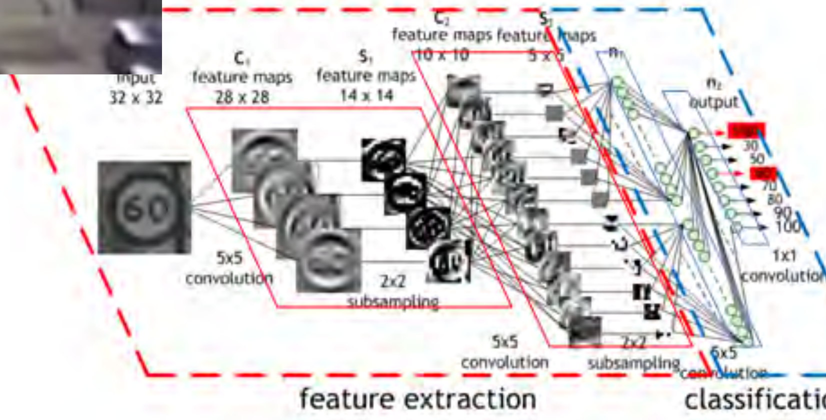
- Multi-layer Neural Net



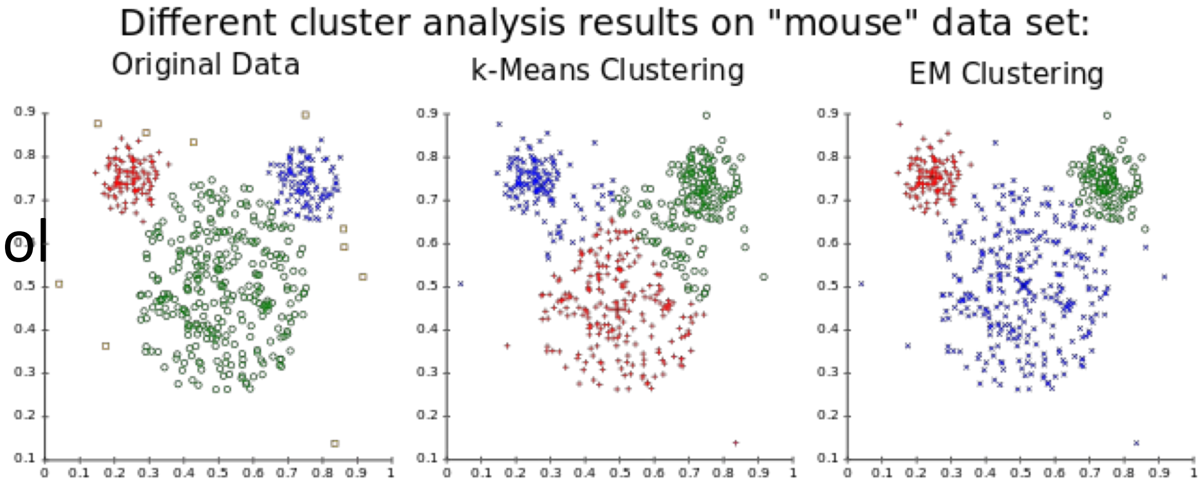
A critica

- Back propagatio
  - Basic idea in 19
    - *In 1968, I prop*
      concept of a b
      flowing back f
  - Reinvented by (
- Use of CNN
  - Application in 1
    - LeCun, Y; Bose
      Hubbard, W; Jackel, L, "Backpropagation Applied to
      Handwritten Zip Code Recognition," in Neural
      Computation , vol.1, no.4, pp.541-551, Dec. 1989

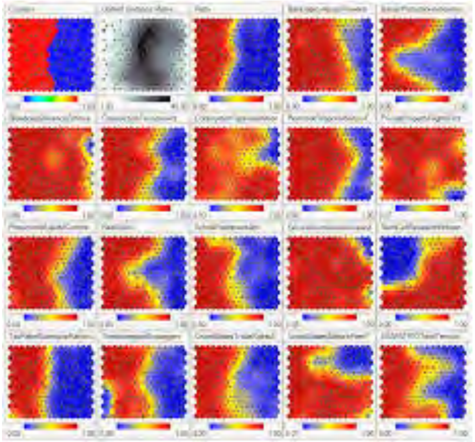




A critical history of deep learning



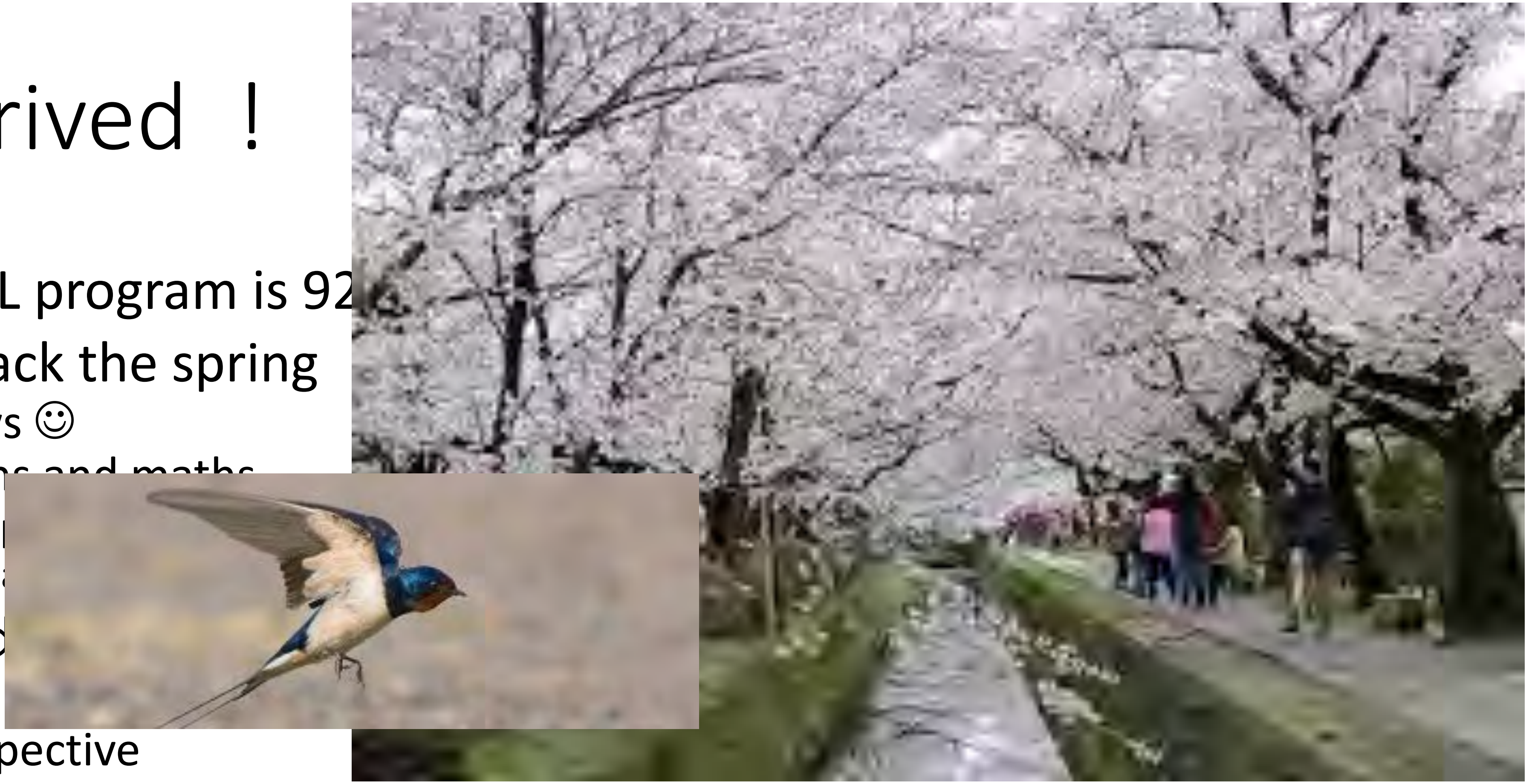
- Neural Nets Go Unsupervised
  - Using NN as a universal compression tool
  - Clustering
    - Kohonen maps
- Fusioning information
  - Belief propagation networks
    - Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines*. Cognitive science, 9(1), 147-169.
  - Graphical models



A critical h

- The glacial age
  - Historically, th
    neural netwo
    very strong b
    Conference o
    should not ac
    was not appr
    no papers wit
    papers about neural networks. That was only a few years ago. And one of the IEEE journals actually had an official policy of [not accepting your papers]. So, it was a strong belief.”

Spring arrived !



- This year ICML program is 92
- What bring back the spring
  - Not swallows ☺
  - Maths, maths and maths
    - Convex Op
    - Differenti
- But the ingred
  - Amateurism
  - Lack of perspective
  - Generalization issue
  - Non reproducibility
  - Boredness
  - Hegemony

So what are the area of interest ?

- Automatic misconfiguration detection
- Data mining to extract new attacks
- New feature extraction tools
  - Highly Non linear
  - Information fusion
    - Heterogeneous source
- High speed computation

Large-scale graph monitoring

An application to overall monitoring of Internet through BGP feeds

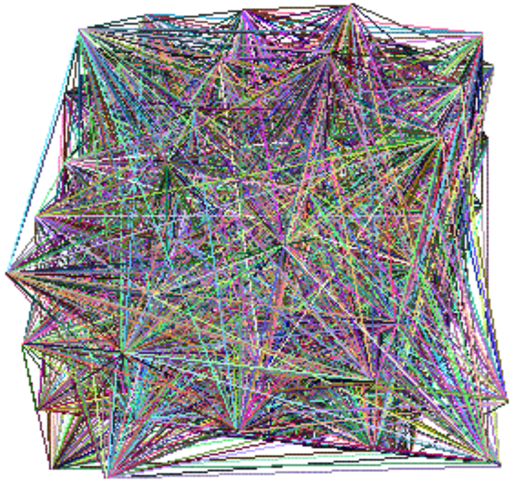
Kavé Salamatian,

Professor of Computer Science, University of Savoie

Distinguished visitor professor in Chinese Academy of Science

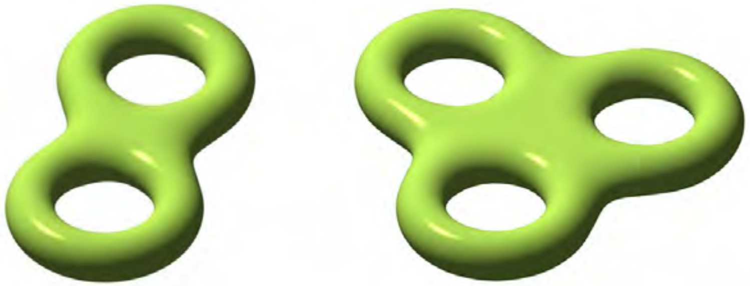
Holder of Presidential Award of the Chinese Academy of Science

Large graphs monitoring

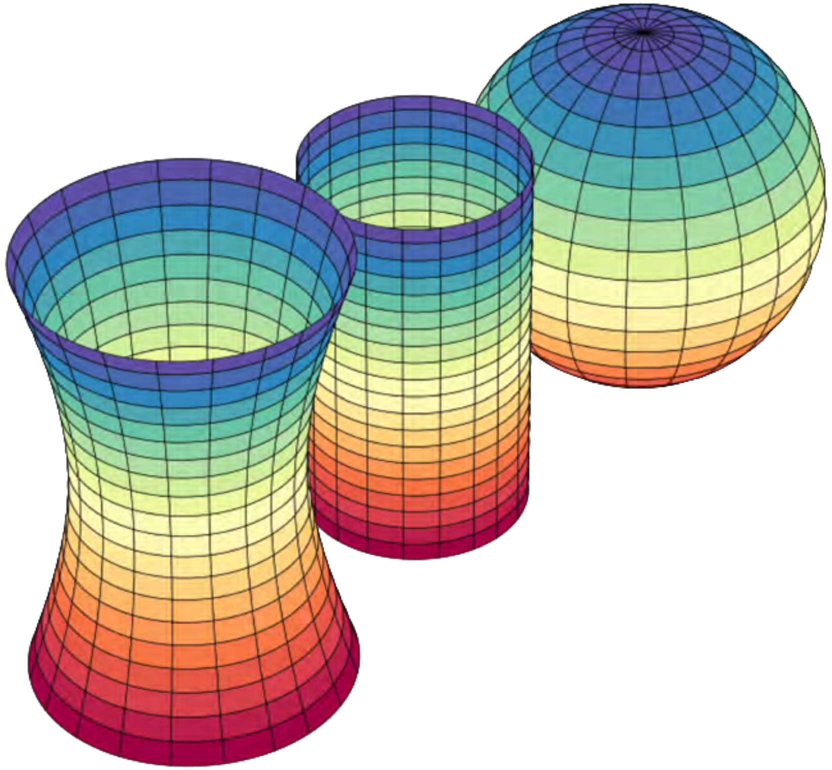


- Graphs are complex object
  - Nodes and links represent things that are of different nature
  - All change to graph are local but some of them have global effe
- Graph monitoring is the process of deciding if a local change wiIl lead to global changes or not ?
  - Large set of applications
    - Computer Networks, biology, social networks
- How can we know that a local change is scaling into global?

On geometry and topology



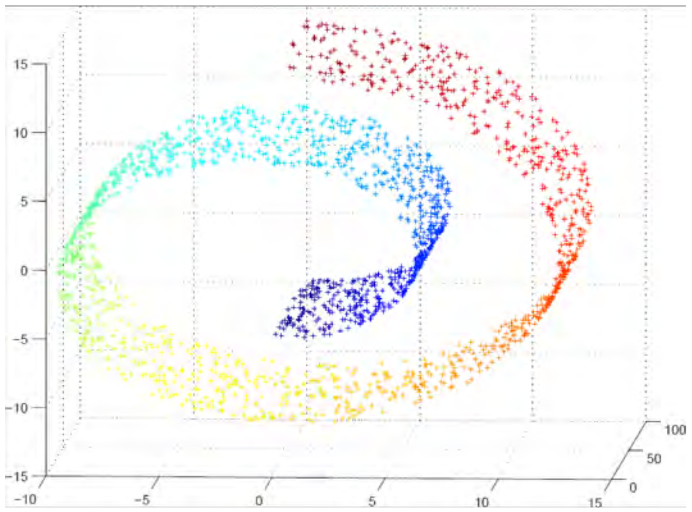
- In 18th century Gauss raised this question:
  - Do an ant moving on a shape can know what is the shape?
- Curvature
  - Describe how geodesics converge or diverge
- Gauss-Bonnet theorem

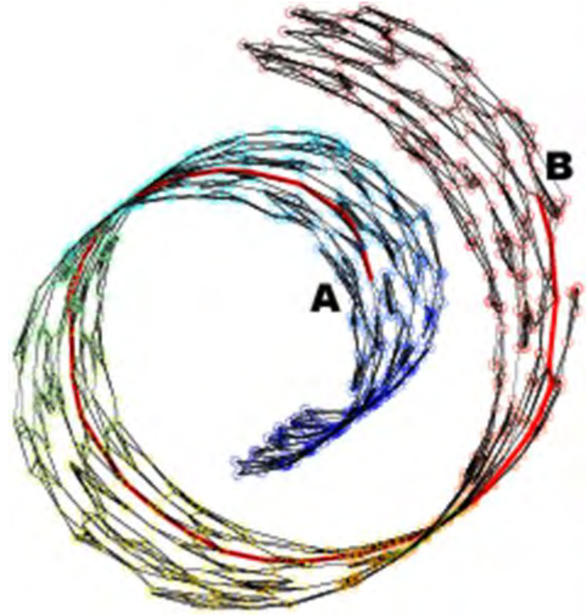


$$\int_M K\,dA + \int_{\partial M} k_g\,ds = 2\pi\chi(M),$$

Extension to graphs ?

- Graph embeddings
  - Transpose a graph into the manifold where an ε-distance linking will create the graph
  - Too complex
    - The manifold was more complex than the graph
- Can we just reproduce curvature structure ?
  - Forman Curvature
  - Ricci-Ollivier Curvature
  - …





Optimal transport

- Evaluate the cost of transferring some distribution of mass over nodes of a graph to another
  - Consider distribution of mass $\mu(x)$ and $\nu(x)$ over all nodes x in the graph
    - $\sum \mu(x) = 1$ and $\sum \nu(x) = 1$
  - Optimal transport is $\theta^*(\mu, \nu) = \arg\min_{\theta} \sum_{x, u \in V} \theta(x, y) d(x, y),$ where d(x,y)

is the cost of transport one unit of mass from x to y and $\theta(x, y)$ is the amount of mass to transport, with constraints

$$\sum_{y \in V} \theta(x, y) = \mu(x) \quad \text{for all } x \in V \qquad \sum_{x \in V} \theta(x, y) = \nu(y) \quad \text{for all } y \in V.$$

  - Boils down to shortest path if all mass is concentrated over two points

- Transportation distance $C(\theta^*, \mu, \nu) \triangleq \sum_{x, y \in V} \theta^*(\mu, \nu) d(x, y)$

Ollivier-Ricci Curvature

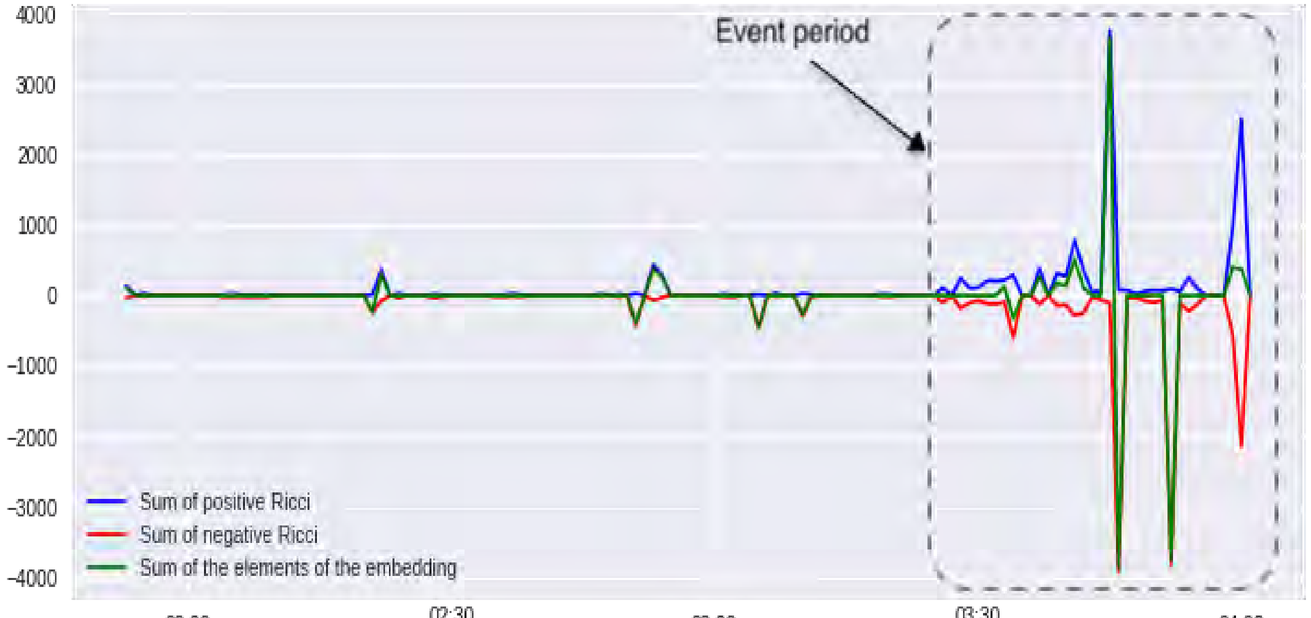
- Optimal transport over a distribution defined over the neighbors of source and destination

$$\kappa(x, y) = 1 - \frac{C(\theta^*, \mu_x, \mu_y)}{d(x, y)},$$

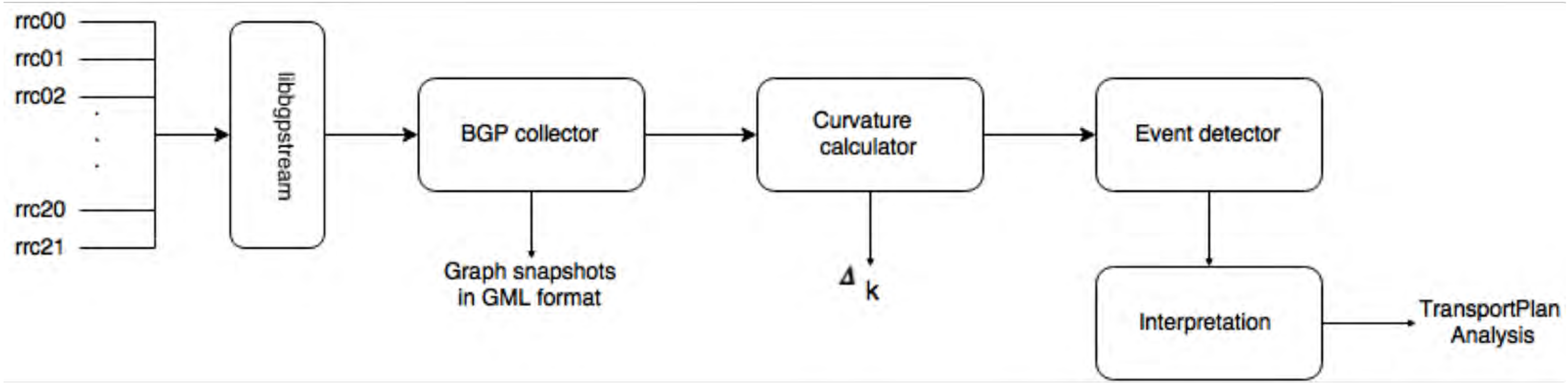
- Examples
  - Over a clique with N nodes the ORC is 1- 1/(N-1).
  - Over an alignment of links the ORC is 0
  - Two star connected by a link ORC is -1+3/2N

Ollivier-Ricci Curvature monitoring system

- Compare the Ollivier- Ricci between all nodes of two snapshots of the graph
- Evaluate the importance of the change by the magnitude of the change
- Does Gauss-Bonnet theorem is valid ?
  - Almost but not in general



BGP monitoring system



Monitoring platform

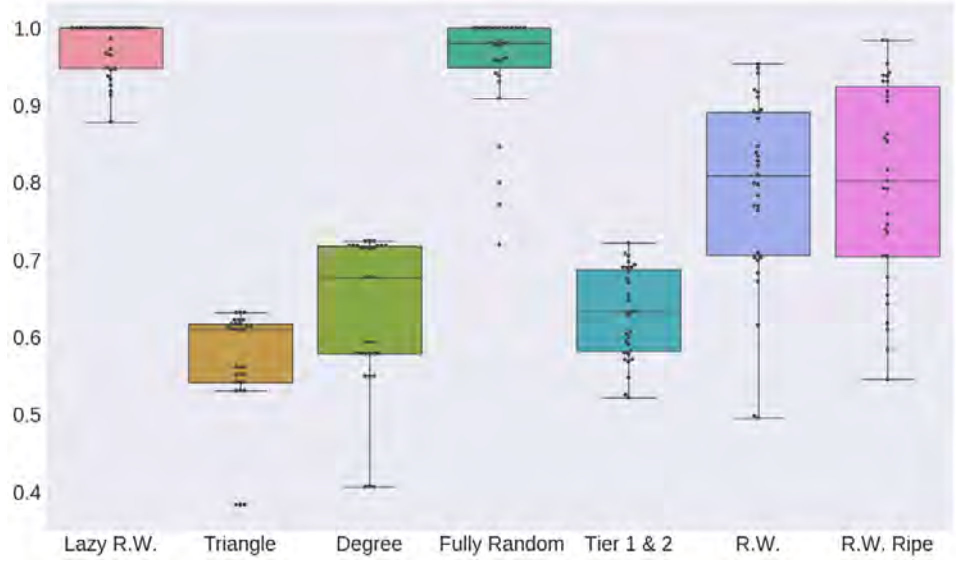
- JSON updates
  - collector': 'rrc19', 'message': 'announce', 'peer': {'address': '197.157.79.173', 'asn': 37271}, 'time': 1515110408, 'fields': {'asPath': ['37271', '6939', '52320', '23106', '23106', '23106', '262700'], 'prefix': '187.102.120.0/21', 'nextHop': '197.157.79.173’},
- Augmented
  - 'flags': {'version': 'v4', 'shortPath': ['37271', '6939', '52320', '23106', '262700'], 'geoPath': ['ZA', 'US', 'CO', 'BR', 'BR'], 'names': ['Workonline Communications(Pty) Ltd', 'Hurricane Electric, Inc.', 'GlobeNet Cabos Submarinos Colombia, S.A.S.', 'Cemig Telecomunicações SA', 'Efibra Telecom LTDA - EPP'], 'risk': 9.262460855949895e-05, 'previousPath': None, 'activePath': None, 'category': None}}
- Each mins one snapshot of the AS level Graph

Landmark selection

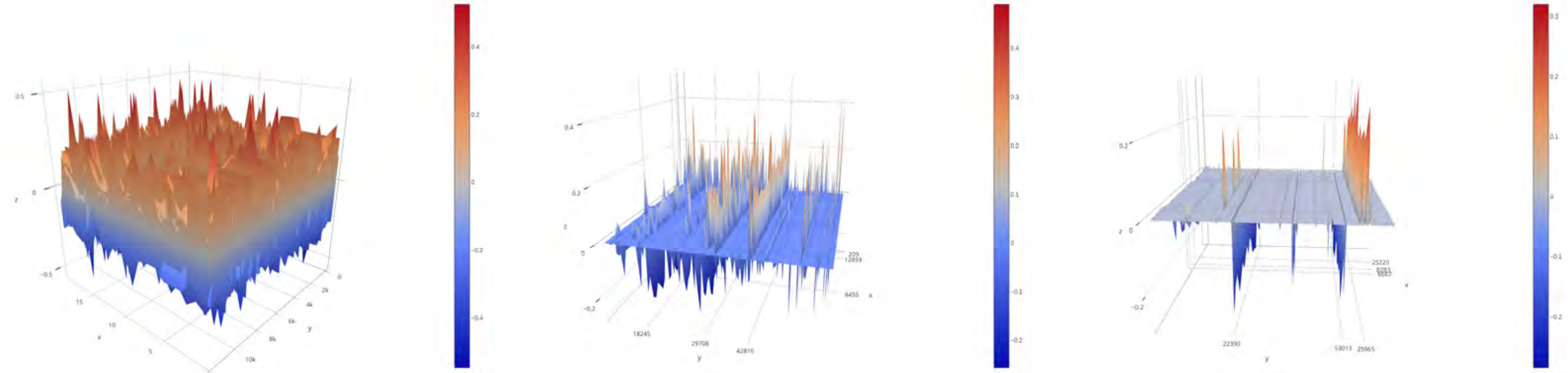
- On a 60k network one cannot afford to calculate the curvature between all nodes
  - We limit this to a set of landmarks and only node that have seen an update in a time window
- Landmarks ?
  - Nodes that are well connected to other nodes but are not close to each other

$$S_1(R) = \frac{\left|\bigcup_{v \in R} N(v)\right|}{\sum_{w \in R} |N(w)|}$$

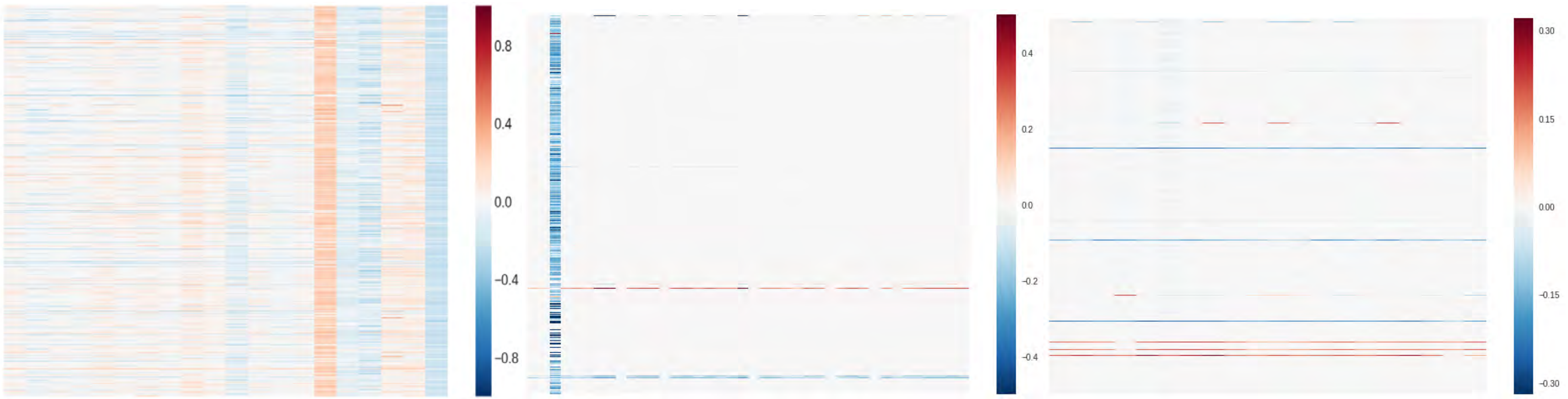
$$S_2(R) = \frac{1}{2|R|} \sum_{v \in R} \sum_{w \in R} d(v, w)$$



Monitoring: comparing curvatures



(a) Global effect event (b) Local effect event (c) Natural drift



Anomaly detector

- Frobenius norm of a matrix

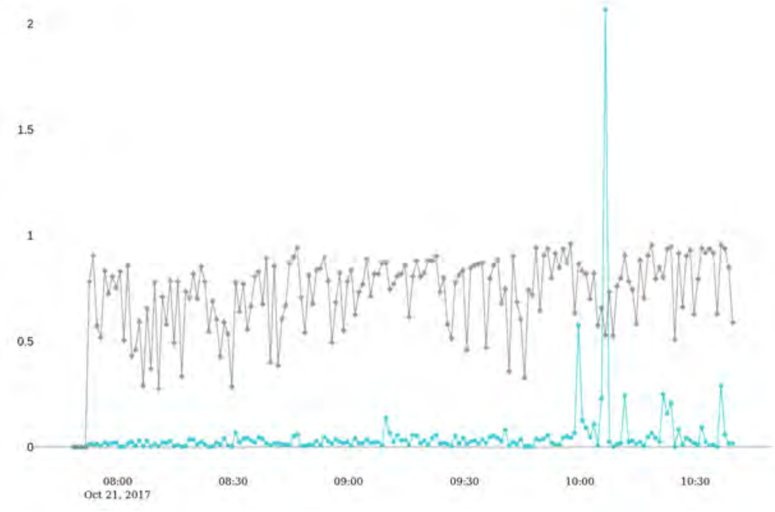
$$\|\Delta^k\|_F = \sum_i \sum_j \left(\delta_{ij}^k\right)^2 .$$

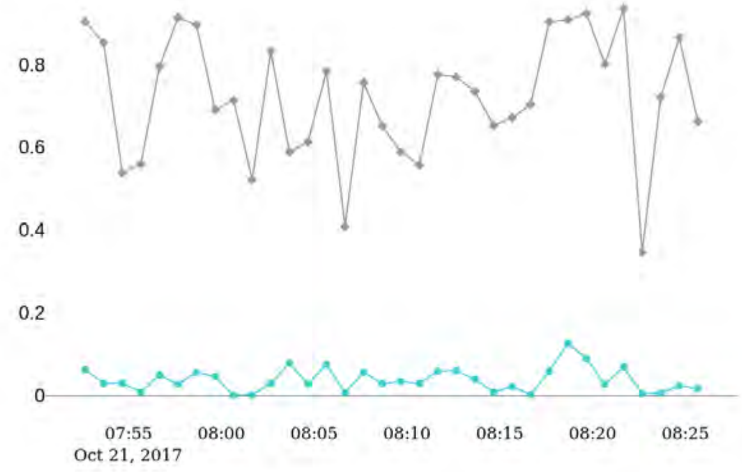
$$\lambda_k^0 = \max_{\|X\|_2 \neq 0} \frac{\|\Delta^k X\|_2}{\|X\|_2} .$$

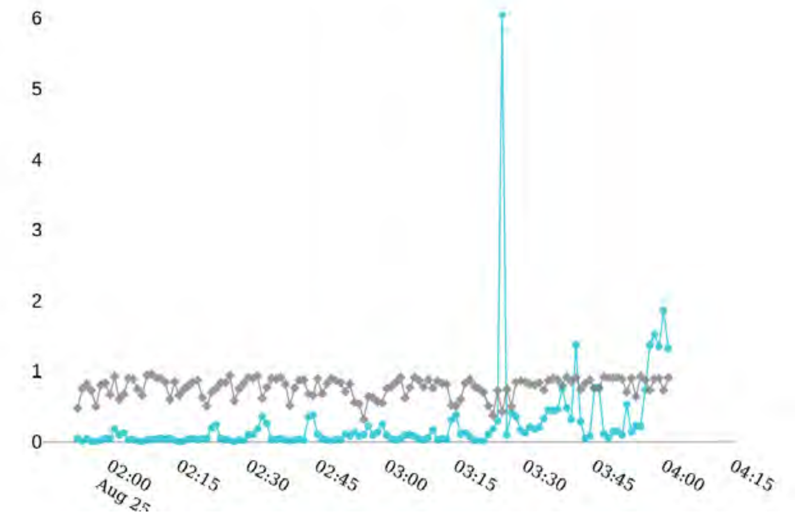
- Largest eigenvalue of the matrix

- We monitor for each difference matrix the Frobenius norm and the Stable rank
  - A large scale anomaly will have A large Frobenus norm and large stable rank

Anomaly detector in the wild

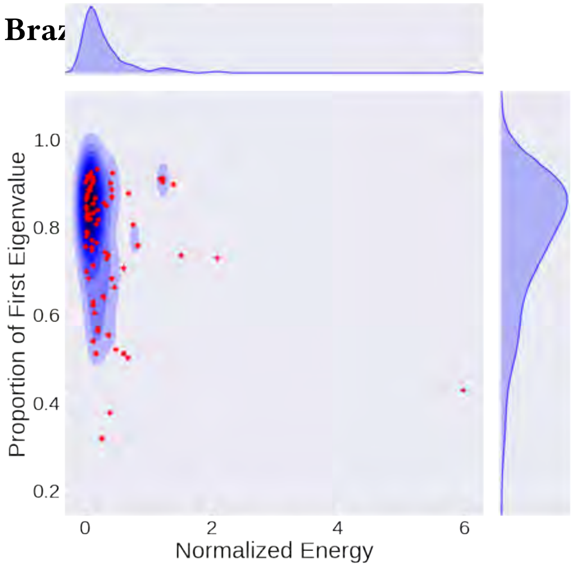




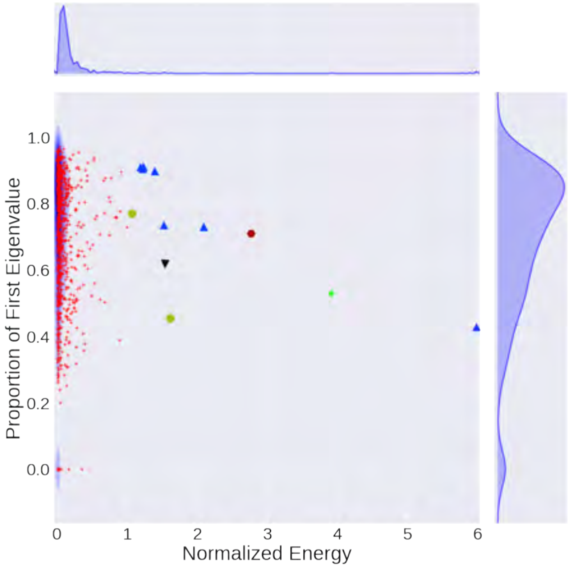


(a) Braz

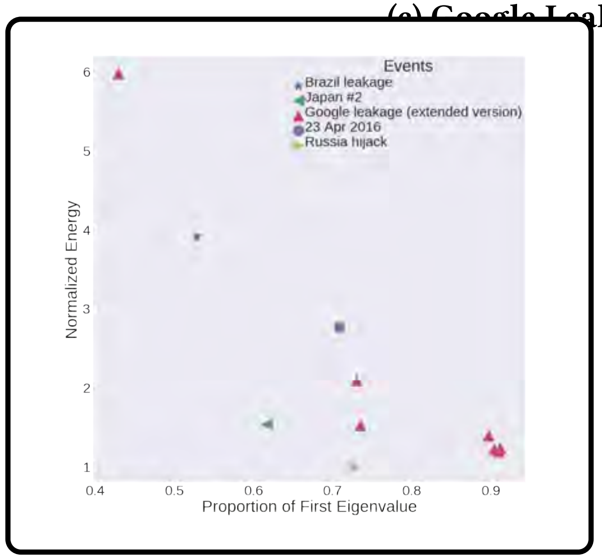
(c) Google Leakage



(a) Phase diagram of the Google leakage

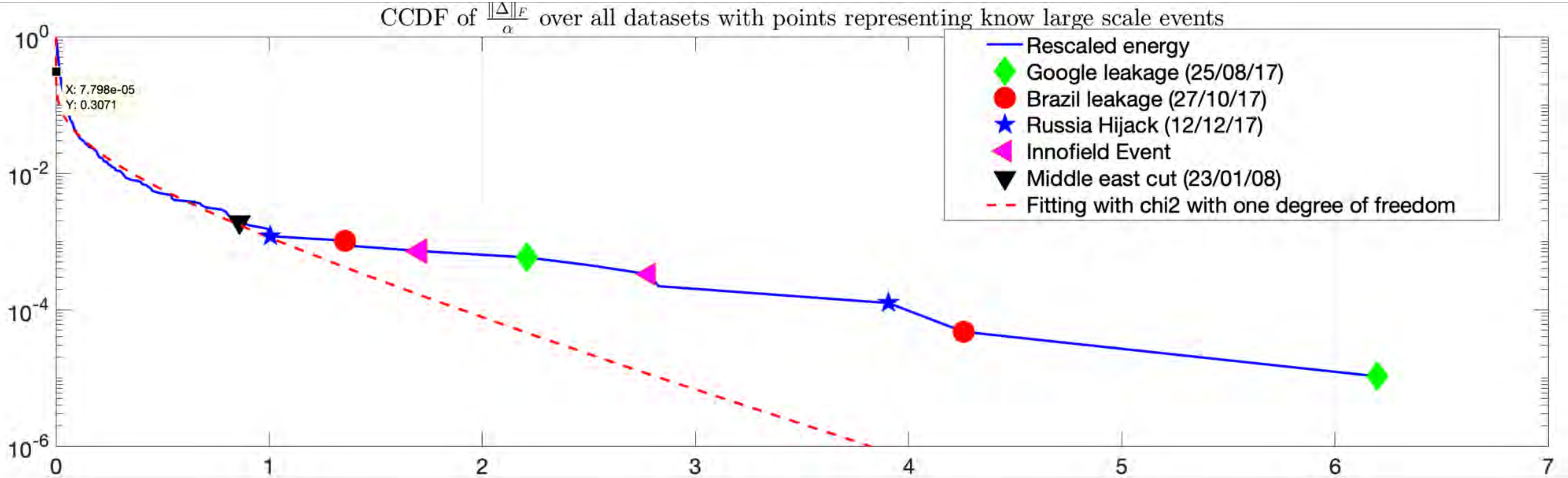


(b) Phase diagram of all datasets



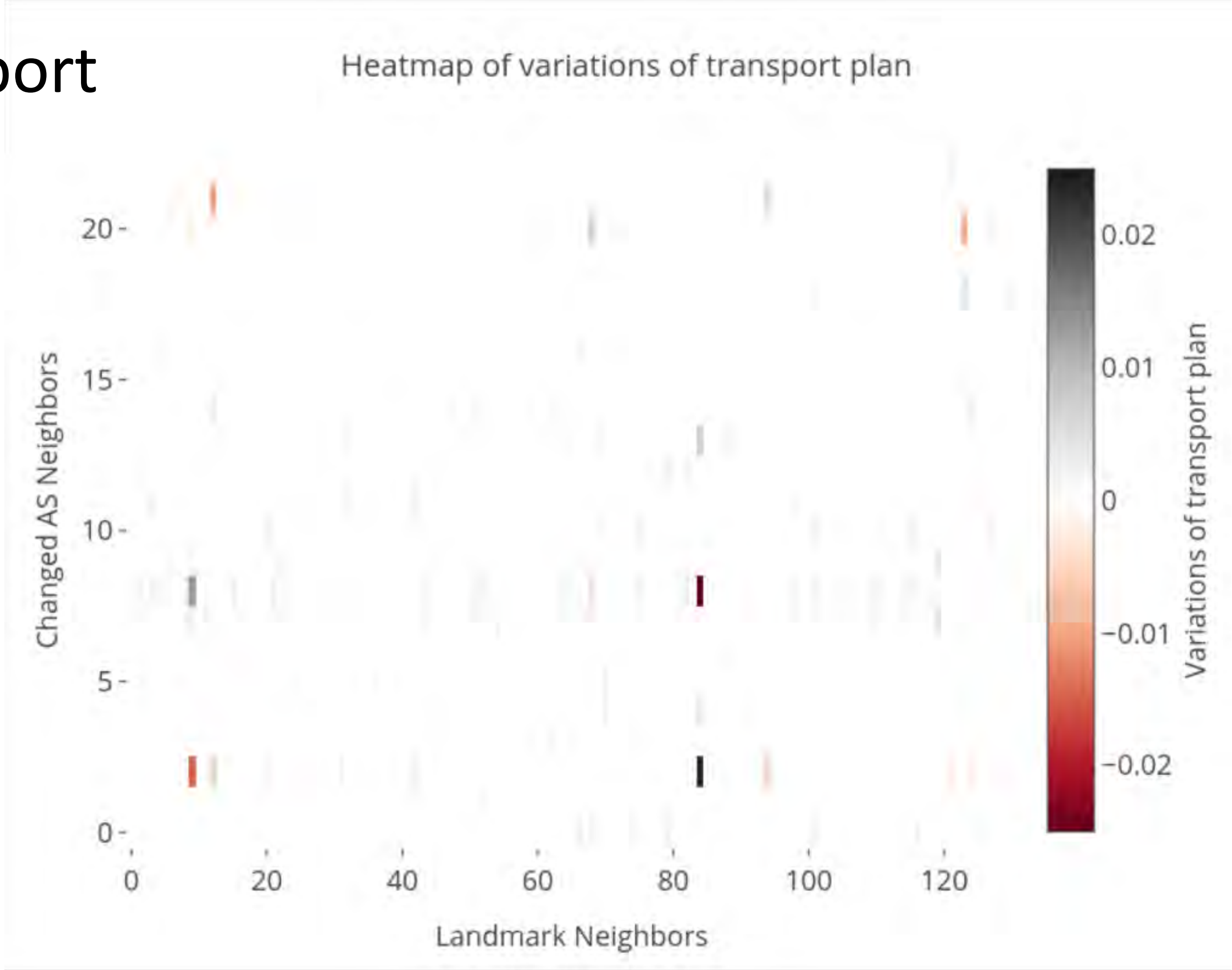
(c) Zoom of the all datasets phase diagram

Calibrating detector



Interpretation of the anomaly detection

- We can use the optimal transport plan changes



Big data challenges

- Processing large graphs
  - Graph are up to 80 k nodes
- Even if the optimal transport is a linear programming we have to solve 1000th of them
  - Distance matrix are node dense
- Computing cluster is needed