Machine Learning in Anomaly Detection

Complex Networks

Radek Mařík

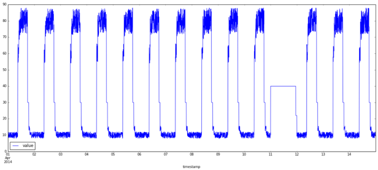
Czech Technical University
Faculty of Electrical Engineering
Department of Telecommunication Engineering
Prague CZ

UDRC Summer School 2017, University of Surrey, UK

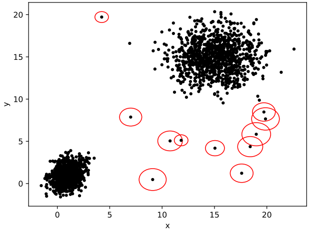
28 June 2017

Anomalies? Outliers?

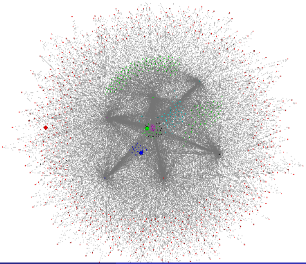
Anomaly



Clouds of points (multi-dimensional)



Complex Network



Outline

What is Anomaly?

Anomaly Definition [CBK09]

Anomaly is a pattern in the data that does not conform to the expected behavior.

- Anomalies are often related to significant real life entities
  - Cyber/Network intrusions, Image Processing / Video surveillance
  - Insurance/Credit card fraud
  - Industrial damage
  - Novel topic in text mining, Customer segmentation
- When an anomaly occurs, its consequences can be quite dramatic and quite often in a negative sense.

Why do we try to detect anomalies?

- Prevention (crime, device failure, production optimization, etc.)
- Novelty detection (technology trends, opinion)
- Knowledge extension (differences from known principles, laws)

Anomaly Basic Characteristics

Observations

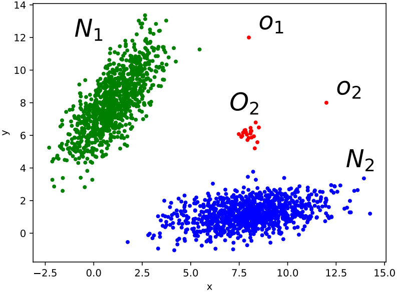
- Pattern . . . repeated
- Expected . . . "expected value" - extremal values, rare occurrences
- Expected behavior . . . subsets, temporally repeated
- Conform . . . simillarity, difference, distance, measurable

An anomaly as a data object (or a group of objects) that is [ATK15]

- **Rare** . . . e.g. a rare combination of categorical attribute values,
- **Isolated** . . . e.g. far-away point in n-dimensional space,
- **Surprising** . . . e.g. data instances do not fit well in a mental/statistical model.
- It requires too **many bits** to describe under the Minimum Description Length (MDL) principle [Ris78, Gru05].

Simple Example - Multidimensional Space [CBK09]

- N_1 and N_2 are regions of "normal" behavior
- Points o_1 and o_2 are anomalies
- Points in region O_3 are anomalies



Normal behavior

- **Normal distribution** . . . $N(\mu, \sigma)$.
  Further, it will be referred as Gaussian distribution
- **Normal behavior/pattern** . . . it is expected, not anomalous.

Key Challenges [CBK09]

- Defining a representative normal region is challenging
  - The *boundary* between normal and anomalous behavior is often not precise
  - The exact notion of an outlier is different for different *application domains*
- Availability of labeled data for training/validation
- Data might contain noise
  - Normal data - noise - anomaly
- Normal behavior keeps evolving, i.e. it is not static

Anomaly Types [CBK09]

- **Point/Global** Anomalies
  - An individual data instance is anomalous w.r.t. the data
- **Contextual** Anomalies
  - An individual data instance is anomalous within a context
  - Requires a notion of *context*
  - Also referred to as conditional anomalies [SWJR07]
- **Collective** Anomalies
  - A collection of related data instances is anomalous
  - The individual instances within a collective anomaly are not anomalous by themselves
  - Requires a relationship among data instances
    - Sequential Data
    - Spatial Data
    - Graph Data
- Online Anomaly Detection
- Distributed Anomaly Detection

Point Anomaly Types Taxonomy [CBK09]

- Classification based
  - Rule based
  - Bayesian Networks based
  - Neural Networks based
  - SVM based
- Nearest Neighbor Based
  - With respect to each instance local neighborhood
  - Density Based
    - Local Outlier Factor (LOF)
    - Connectivity-based Outlier Factor (COF)
  - Distance Based
- Clustering Based
  - With respect to the cluster each instance belongs to
- Statistical
  - Parametric
    - Gaussian model
    - Regression model
  - Non-parametric
    - Histogram based
    - Kernel function based
- Others
  - Information Theory
  - Spectral Decomposition
  - Visualization Based

Input Data [Agg17]

- Record data
  - Univariate
  - Multivariate
- Attributes
  - Binary/Boolean
  - Categorical
  - Continuous
  - Hybrid
- Relations
  - Sequential
    - Temporal
  - Spatial
  - Spatio-temporal
  - **Long range correlations**
  - **Graph**
- Data Quality
  - Data Fusion
  - Data Cleansing
  - Consistency maintenance
- Processing
  - Online/Offline processing
  - Distributed processing
  - **Analysis** × **Production**
    - Feature/Property searching/selection
    - Selected features detection
- Data Volume
  - Dense/**Sparse**
  - Low/High dimensions
  - Low/**Large volumes**
  - Big data
  - Internet of Things

Input Data - The Old Kingdom of Egypt [MD15]



- Continuous . . . tomb dimensions
- Categorical . . . titles
- Binary, boolean . . . titles
- Multivariate . . . people, titles, tombs
- Temporal . . . dynasties, king reigns
- Spatio-temporal . . . location of tombs in time

Output of Anomaly Detection [CBK09]

- **Score**
  - Each test instance is assigned an anomaly score
  - E.g. Euclidean, Mahalanobis, Frobenius norms
  - Allows the output to be ranked
  - Requires an additional threshold parameter
- **Label**
  - Each test instance is given a normal or anomaly label
  - This is especially true of classification-based approaches

Data Supervision [HA04]

Availability of class labels reflected by data processing

- **Supervised** Anomaly Detection
  - Labels available for both normal data and anomalies
    - $\Longrightarrow$ machine learning
    - $\Longrightarrow$ the classification problem is often **highly imbalanced**
- **Semi-supervised** Anomaly Detection
  1. Labels available only for **normal** data
     - $\Longrightarrow$ one-class learning, **thresholding**
  2. Labels available only for **anomalous** data
     - $\Longrightarrow$ one-class learning
     - $\Longrightarrow$ highly tuned commercial tools for network monitoring and analysis
- **Unsupervised** Anomaly Detection
  - No labels assumed
  - Often based on the assumption that anomalies are very **rare** compared to normal data
    - $\Longrightarrow$ clustering

What is an outlier? [KKZ10, Agg17]

Hawkins (1980) [Haw80]

An **outlier** is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.

Barnett and Lewis (1994) [BL94]

An outlying observation, or **outlier**, is one that appears to deviate markedly from other members of the sample in which it occurs.

- Statistics-based intuition
- Normal data objects follow a "generating mechanism", e.g. some given statistical process
- Abnormal objects deviate from this generating mechanism

Grubbs's Test [Gru50, Gru69]

- To detect **a single outlier in a univariate data** set that follows an approximately **Normal (Gaussian) distribution**
- Also known as the maximum normed residual test
- The test considers the maximum absolute difference between observations $x_i$and the mean $\bar{x}$, normalised with respect to the sample standard deviation s:

$$G_{\max} = \max_i \left|\frac{x_i - \bar{x}}{s}\right|$$

- And considers the chance of such an extreme value occurring given the number of observations, given that the data are Normally distributed.
  - H_0: The observation is not different than the sample population.
  - H_a: The observation is different than the sample population.

Test statistics

- Significance level α
- Critical region: for the two-sided test, the hypothesis of no outliers is rejected if

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{(t_\alpha/(2N), N-2)^2}{N-2+(t_\alpha/(2N), N-2)^2}}$$

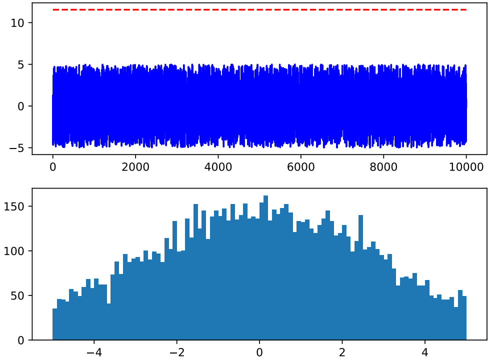
with $t_\alpha/(2N), N-2$ denoting the critical value of the t distribution with $(N-2)$ degrees of freedom and a significance level of $\alpha/(2N)$.

Grubbs's Test - Example [Gru50, Gru69, HA04]

Example 1

- $n = 93, \bar{x} = 52.2, s = 1.38, G_{\max} = (57.0 - 52.2)/1.38 = 3.49$
- From the G table at $n = 93$ and $\alpha = 0.05$ the critical value is 3.18.
- Since $3.49 > 3.18$, reject H_0.
- The observation is from a different population $(G_{3.49}, p < 0.025)$.

- The test gives false results for some especially **asymmetric** distributions.
- A significant **gap** between the data produced by the normal bounded generating process and the level at which observations are treated as outliers.
- $N(0, 3)$ truncated by $[-5, 5]$
- Significance level $\alpha = 0.05$
- An outlier detected above 11.55



Thresholds based Extreme Value Theory [Mar17b]

- Let $X_1, \ldots, X_N$ be a (discrete) sequence of independent random variables having a common distribution function F that is unknown.
- A given volume (**block**) of n observations to which we relate occurrences of extremal values
- The distribution of block maxima is created.
- We set a threshold τ as a robust estimate of the upper bound of signal values

$$\tau = p_{0.975,n} + \mathsf{iqr}_n$$

  - where n is a sufficiently large block size depending on the application problem and its setup,
  - percentile $p_{0.975,n}$,
  - iqr_n is the interquartile range used instead of the standard deviation.

$$\mathsf{iqr}_n = p_{0.75,n} - p_{0.25,n}$$

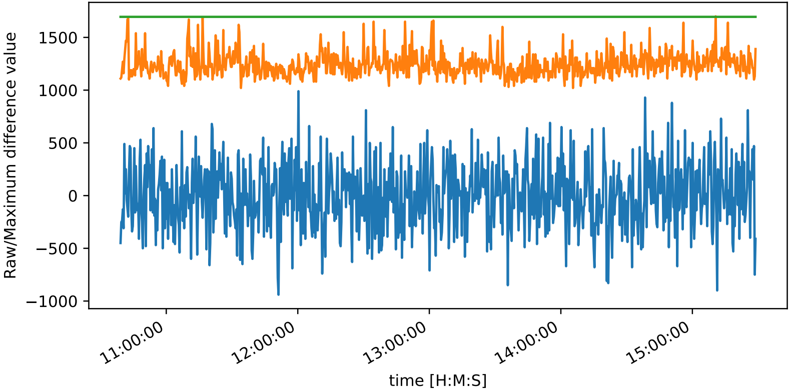
Extreme Value Theory (EVT) [Col01]

- $\{X_1, X_2, X_3, \ldots\}$ is a sequence of iid random variables
- The **block maxima** $Z_{n,i} = \max(X_1, \ldots X_n)$, $i = 1, \ldots, m$
- A random variable Z is said to have a **generalised extreme value distribution** (GEV) [Jen55] with scale parameter $\sigma > 0$, location parameter μ and shape parameter ξ, if its cumulative distribution function is

$$G(z) = \exp\{-[1 + \xi(\frac{z-\mu}{\sigma})]^{-1/\xi}\} \tag{1}$$

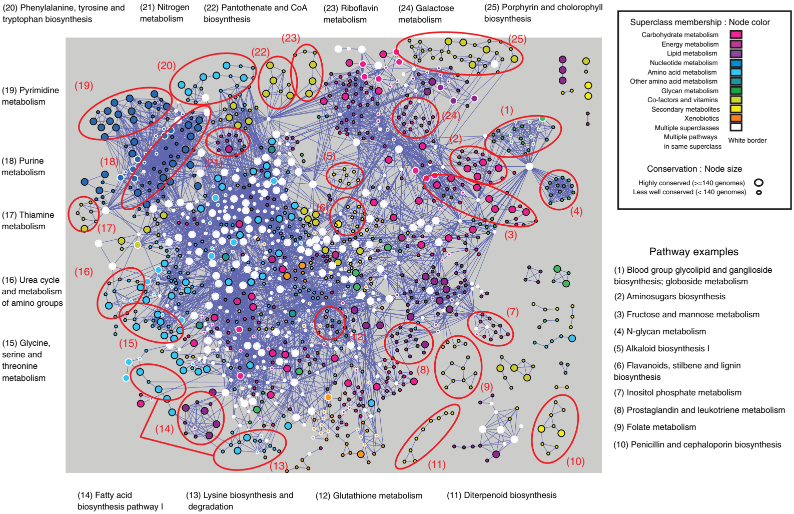
- defined on the set $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$,
- where the parameters satisfy $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \xi < \infty$
- The shape parameter ξ split the GEV family into three subfamilies
  - $\xi > 0$ . . . the **Frechet** family which density decays polynomially and $z_+ = \infty$.
  - The limit of $G(z)$ for $\xi \to 0$ leads the **Gumbel** family which density decays exponentially and $z_+ = \infty$.
  - $\xi < 0$ . . . the **Weibull** family, z_+ is finite, **the threshold**

Example - EVT based Threshold [Col01]

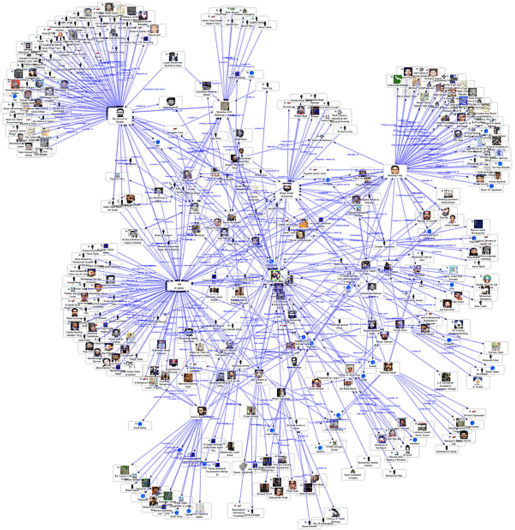


- Linear optical sensor stream data,
- **Blue** - the subsampled raw signal, **Orange** - block maxima,
- **Green** - the threshold

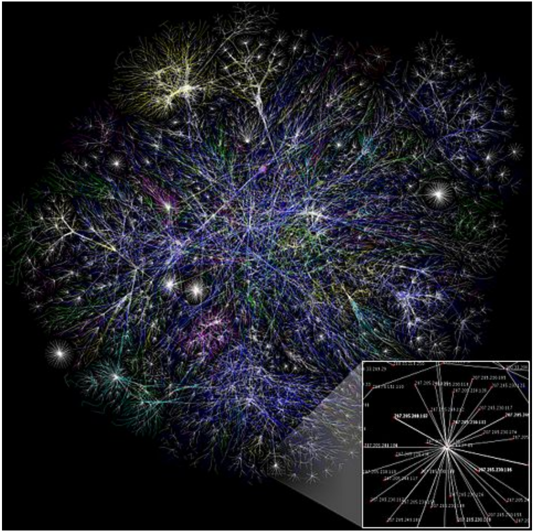
Conservation within the global metabolic network [PASP09]



Link Analysis of the Al Qaeda Terrorist Network [FMS]

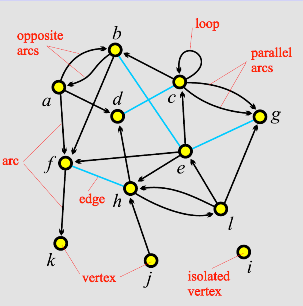


Internet map in 1995 [Bri95]



Graph [Weh13]

A graph is a set of vertices and a set of lines between pairs of vertices.



- **Actor** - vertex, node, point
- **Relation** - line, edge, arc, link, tie
  - Edge = undirected line, $\{c, d\}$
    c and d are end vertices
  - Arc = directed line, (a, d)
    a is the initial vertex, (source, start)
    d is the terminal vertex, (target, end)
  - Parallel (multiple) arcs/edges are only allowed in multigraphs with more than one relation (set of lines).
  - Loop (self-choice)

We focus on simple graphs!

A simple undirected graph has no loops and no parallel edges.
A simple directed graph has no parallel arcs.

Network [Weh13]

Network

A network consists of a graph and additional information on the vertices or the lines of the graph.

Formally, a network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \mathcal{P}, \mathcal{W})$ consists of:

- A graph $\mathcal{G} = (\mathcal{V}, \mathcal{L})$, where
  - $\mathcal{V}$ is the set of vertices,
  - $\mathcal{A}$ is the set of arcs,
  - $\mathcal{E}$ is the set of edges, and
  - $\mathcal{L} = \mathcal{E} \cup \mathcal{A}$ is the set of lines.
- $\mathcal{P}$ vertex value functions / properties: $p : \mathcal{V} \to A$
- $\mathcal{W}$ line value functions / weights: $w : \mathcal{L} \to B$

- **Long range dependencies** vs. multidimensional space
- **Specific topological properties**
- **Large/Huge volumes** of **sparse** data records

Networks Focused on Relations [Weh13]

RELATIONS MATTER!

Contrasted with both an *atomistic* perspective or a *whole-group* perspective

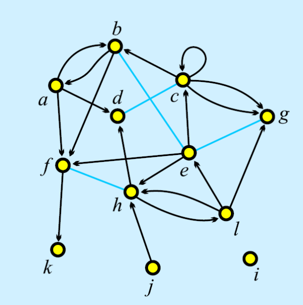
Social Network Analysis (SNA)

- Humanities and social science
- Activities and structures tied with people
  - Shopping basket analysis, targeted advertising
  - Enterprise processes analysis(people cooperation, good distribution)

Complex Network Analysis (CNA)

- Uses the same method as SNA
- Applied to all domains of human acting
- Biology, military, computer network, citations, telecommunication

Vertex Degree [Weh13]



- **Degree** of vertex i,
  $deg(i) = d_i = k_i = \sum_{j=1}^{n} A_{ij}$
  = the number of lines with i as end-vertex,
  (end-vertex is both initial and terminal)
- **Indegree** of vertex i, $indeg(i), deg^{+}(i)$
  $= k_i^{\text{in}} = \sum_{j=1}^{n} A_{ij}$ the number of lines with v as terminal vertex
- **Outdegree** of vertex j, $outdeg(j), deg^{-}(j)$
  $= k_j^{\text{out}} = \sum_{i=1}^{n} A_{ij}$ the number of lines with j as initial vertex.

Example 2

$n = 12,\ m = 23,\ deg^{+}(e) = 3,\ deg^{-}(e) = 5,\ deg(e) = 6$

$$\sum_{v \in \mathcal{V}} deg^{+}(v) = \sum_{v \in \mathcal{V}} deg^{-}(v) = |\mathcal{A}| + 2|\mathcal{E}|$$

Network Fundamental Matrices [New10, EK10]

- The **adjacency matrix** $\mathbf{A}$ of a simple graph is the matrix with element A_{ij} such that

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between vertices } i \text{ and } j, \\ 0 & \text{otherwise} \end{cases}$$

- The adjacency matrix of a directed network has matrix elements

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge from } j \text{ to } i, \\ 0 & \text{otherwise} \end{cases}$$

- The **graph Laplacian** is the matrix

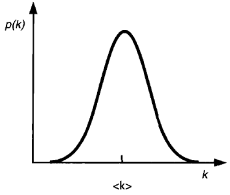
$$\mathbf{L} = \mathbf{D} - \mathbf{A}$$

where

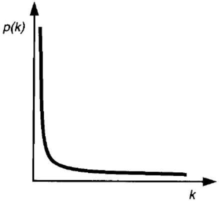
$$\mathbf{D} = \begin{pmatrix} k_1 & 0 & 0 & \cdots \\ 0 & k_2 & 0 & \cdots \\ 0 & 0 & k_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Degree Heterogeneity [Weh13]

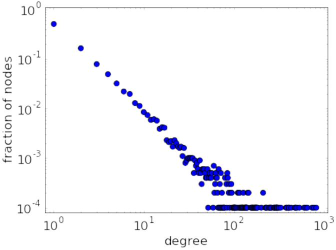
- Not all nodes show the same activity (degree) in networks.
- Some nodes show an astounding activity.
- Degree is most of all a question of tie formation cost.
  - Preferential attachment
  - Fitness model



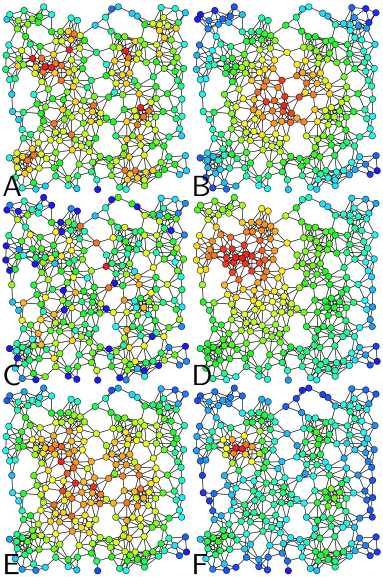
Gaussian



Skewed Distributions



Centrality Measures - Importance of Nodes [Roc12]

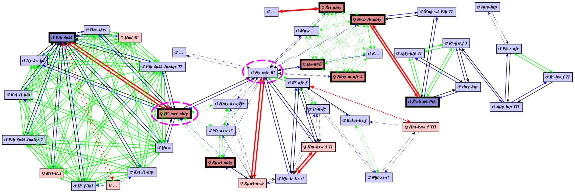


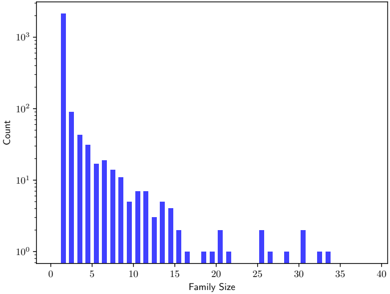
- Low → middle → high values
- **A** Degree centrality,
  - Node Activity
- **B** Closeness centrality,
  - Distance to other nodes
- **C** Betweenness centrality,
  - Intermediate Position
- **D** Eigenvector centrality,
  - Important nodes have important friends
- **E** Katz centrality,
  - The relative influence of a node within a network
- **F** Alpha centrality
  - Important nodes have important friends for asymmetric relations

Egypt Data - Family Formation [DM15]

Ny-wśr-Rꜥ	0.647
Ḫꜥ-mrr-nbty	0.424
Nwb-ỉb-nbty	0.351
Śꜥnḫ-wỉ-Ptḥ	0.290
Rꜥ-ḫw.f I	0.180
Rꜥ-nfr.f	0.139
ꜣḫty-ḥtp III	0.139
Ptḥ-špśś	0.082
Pḥ-r-nfr III	0.048
Šrt-nbty I	0.048

People with
the top 10 highest betweenness





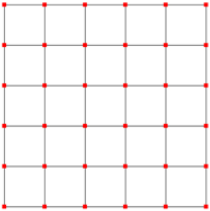
Extended family size distribution

Random Graphs

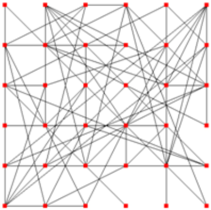
- Basic idea
  - Edges are added at random between a fixed number N of vertices
  - Each instance is a snapshot at a particular time of a stochastic process, starting with unconnected vertices and for every time unit adding a new edge
- Four basic models of complex networks
  - Regular lattices (meshes) and trees
  - Erdös-Renyi Random Graphs (ER)
    - A disconnected set of nodes that are paired with a uniform probability.
  - Watts-Strogatz Models [WS98] (SW)
    - **Small-world networks**
    - Connections between the nodes in a regular graph were rewired with a certain probability
  - Barabási-Albert Model [BAJ99] (SF)
    - **Scale-free networks** characterized by a highly heterogeneous degree distribution, which follows a "power-law"

$$P(k) \sim k^{-\gamma}$$

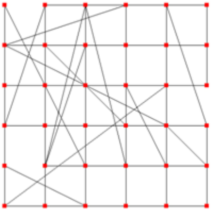
Complex Network Models [GDZ+15]



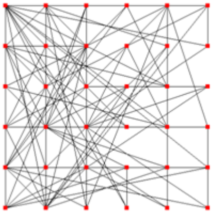
(a) Regular lattice ($p = 0$)



(b) Random network ($p = 1$)



(c) Small-world ($p = 0.01$)



(d) Scale-free ($n_0 = 3$, $m_0 = 3$)

Summary of Approaches [New06, Weh13, CRTV07, HK13]

- The **density** of graph is the proportion of present lines to the maximum possible number of lines.
- **Clustering coefficient** is a measure of the degree to which nodes in a graph tend to cluster together

Global clustering coefficient [HK13]

the ratio of the total number of triangles to the total number of connected triplets.

$$C_g = \frac{2\sum_{i=1}^{N} \ell_i}{\sum_{i=1}^{N} d_i(d_i - 1)}$$

- **Modularity** ... is - up to a normalization constant - the number of edges within communities c minus those for a **null model**
  - *"A good division of a network into communities is not merely one in which there are few edges between communities; it is one in which there are fewer than expected edges between communities".*

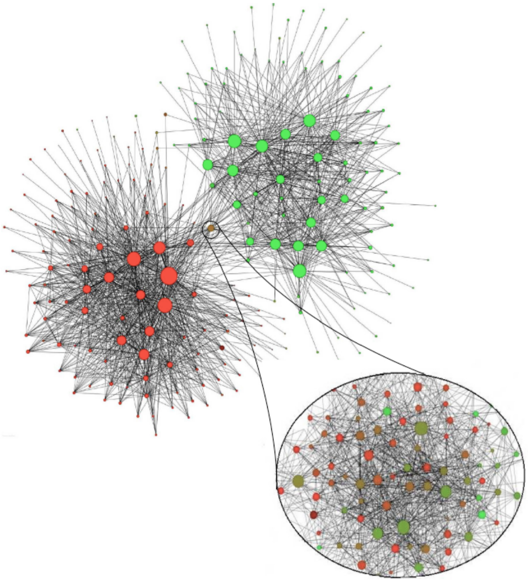
Modularity [New06, BGLL08, New10]

- The **modularity** is, up to a multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random, but with the same node degree distribution.
- A weighted network
- c_i . . . a community of a given node i

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{d_i d_j}{2m} \right] \delta(c_i, c_j)$$

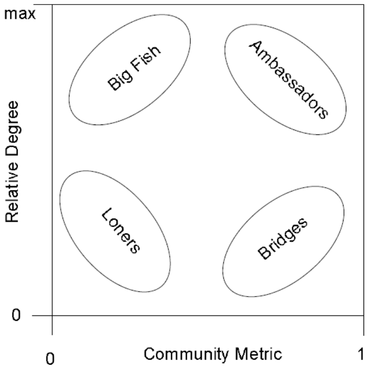
- where
  - A_{ij} . . . weight of the edge between i and j, adjacency matrix
  - $d_i = \sum_j A_{ij}$ . . . degree of i
  - $m = \frac{1}{2} \sum_{i,j} A_{ij}$ . . . total weight
  - $d_i d_j / 2m$ . . . the *expected* number of edges between vertices d_i and d_j
  - $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise
  - $Q \in [-1, 1]$

Belgian Mobile Phone Network - Louvain Method [BGLL08]



- 2.6 millions customers
- Language: Dutch, English, French, German,
- 6.3 millions links
- Weights ... number of call + sms
- Red ... French,
- $> 93\%$ segregated,
- The center ... Brussels

Community-based Node Roles [STE07]



- An *authority* . . . how much knowledge, information, etc. held by a node on a topic.
- A *hub* . . . how well a node 'knows' where to find information on a given topic.
- An *ambassador* has links to many nodes from different communities
- A *big fish* has links only to other nodes in the same community
- A *bridge* . . . serves as bridges between a small number of communities
- A *Loner* . . . a low relative degree and low community score.

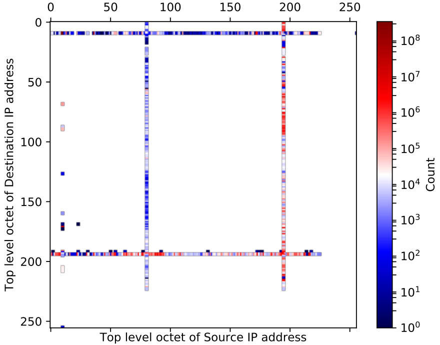
NETFLOW Primary Statistics

- **Netflow**
  - Condensed records on a packet flow
  - Several packets are merged into one netflow record
  - Only 14-20 aggregated metrics

An enterprise traffic as a netflow sample taken during 9 days:

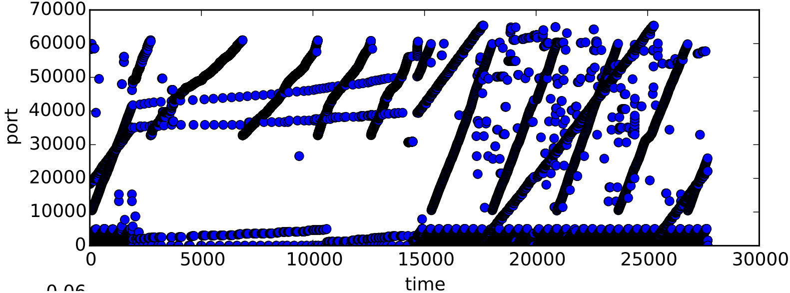
Statistics	Value
Total transported data volume	13,995,690,457,765 $[B]$
Packet count	20,131,367,095
Netflow count	617,326,053
IP address count	686,168
Source IP address count	614,150
Destination IP address count	392,881
Different P2P connections count	2,412,481

Top Level IP Network Projection - Data Sparsity



- Focused on the network of source and destination IP addresses
- Top level octets of IP addresses (160.30.29.17 $\Longrightarrow$ 160)
- A very sparse space
- A rather restricted source domain of IP addresses (as expected)

Port Scanning from xxx.xxx.18.120 - Logical Time Progress



- 617,326,053 netflows $\approx$ 60,000 samples $\times$ sample size 10.000
- $\Longrightarrow$ 60,000 samples might be still visualized with difficulties
- $\Longrightarrow$ 1.000 events can be easily missed with 10,000 sample size

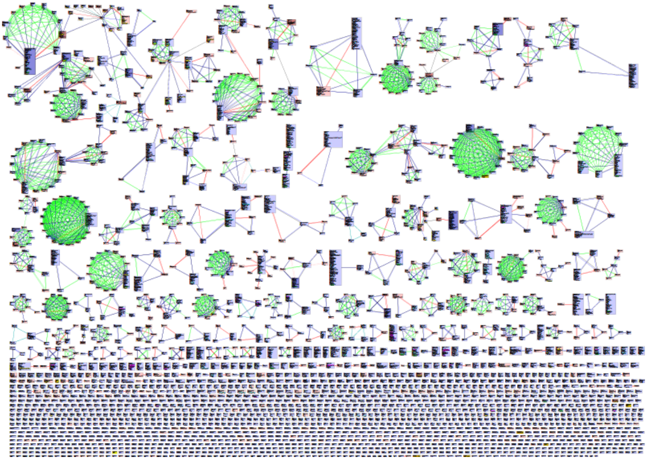
Masters of Social Network Analysis [RP13, Weh13]

- US National Security Agency
- Maintains large programs in social network analysis
- Believed to process 2×10^{10} node and tie updating events per day
- Result:
  "Better Person Centric Analysis"

Types

- **94 entity/node** types
  (*phone numbers*, *e-mail addresses*, *IP addresses*, etc.)
- **164 relationship** types to build "community of interest" profiles
  (*travelsWith*, *hasFather*, *sentForumMessage*, *employs*, etc.)

Egypt Data - Family Recognition

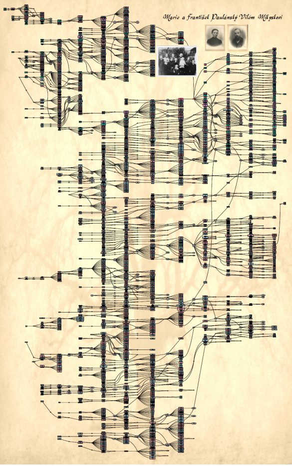


circular layout (yEd)

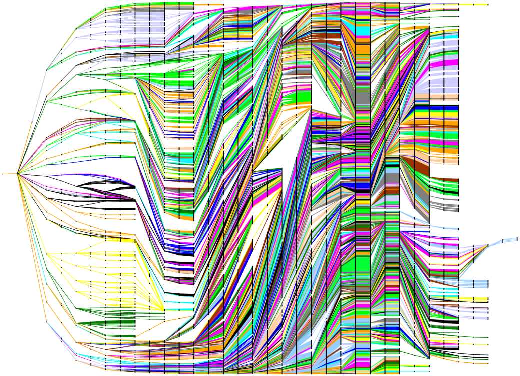
A family:

- Using family designation
  - husband, wife, son, etc.
- A connected graph component
- Sparse data assumed
- Transformed into family tree using marriage nodes

Family Trees[Mar17a]

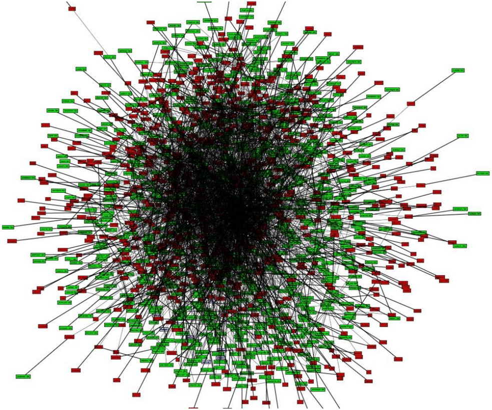


multitree-like tree driven layout, Graphviz



- Taxonomic information ITIS on plants, animals, fungi, and microbes,
- A phylogenetic tree with 945.352 nodes
- **multitree-like tree driven layout**

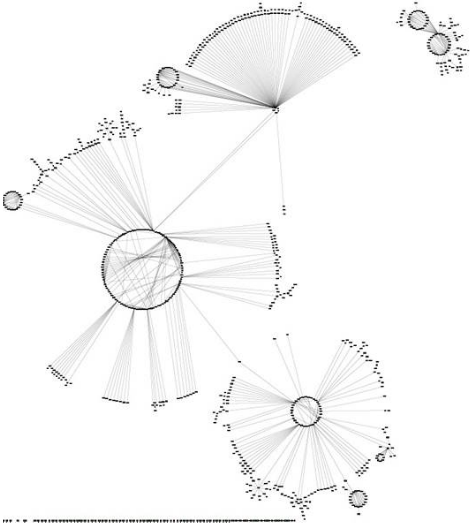
Dependancy of External Symbols in Mainframe Assembly Software



Fruchterman-Reingold force-driven layout

- A software product . . . over 10.000.000 lines of code
- Over 400 modules . . . red
- External symbols . . . green
- Thick line . . . the definition of a symbol
- Thin line . . . a reference to a symbol
-
- *Where should the developer start with a bug analysis?*

Assembly Software - Recovered Architecture



double-circular layout - yEd

Approach to Complex Networks

- One needs to distinguish between **analysis** and **production** phases
- Some phenomena appear only with sufficiently large data volumes (emergent events)
- Volume
  - A number of suitable tools ... HDF5, ElasticSearch, Clouds
  - Capable to operate with terabytes of data
- Visualization
  - Critical if anomaly features are not known
  - At present, there is no obvious choice of a tool and a network layout given a particular problem.
  - Tools do not often scale with data volumes (> 10.000 nodes, 10^5 edges)
  - GGobi, Pajek, NetworkX, SNAP, Tulip, Gephi, Cytospace, yEd, D3.js
  - Aspects: data volume, interactions with the user

Network Anomalies [ATK15]

Networks

- Structured graph data
- Features/properties attached to nodes and edges
- Long-range correlations (data objects exhibit inter-dependencies)
- Specific topological patterns

Various settings of a general framework

- Unsupervised vs. (semi-)supervised approaches,
- Statis vs. dynamic graphs,
- Attributed vs. plain graphs.
- Effectiveness, scalability, generality, and robustness aspects.
- Class imbalance and asymmetric error (rare abnormal $\times$ normal)
- Root cause analysis: Why is it anomalous?

Anomalies in Static Plain Graphs [ATK15]

- **Structure** based methods
  - **Feature** based approaches
    - *Extract features and used outlier detection graph-centric features*
    - Node-level features (centralities, local clustering coefficient)
    - Node-group-level features (compactness, density, modularity)
    - Global measures (number of connected components, principal eigenvalue)
    - Egonet (1-step neighborhood around a node) (OddBall)
  - **Proximity** based approaches
    - *Measure closeness/proximity of objects*
    - Importance of nodes (PageRank, Personalized PageRank, SimRank)
- **Community** based methods
  - *Search for "bridge" nodes/edges that do not directly belong to any community*
    1. How to find the community of a given node?
    2. How to quantify the level of the given node to be a bridge node?
  - Matrix Factorization
    - Boolean/Binary Matrix Factorization (BMF)
    - Non-negative Matrix Factorization (NMF)

Anomalies in Static Attributed graphs [ATK15]

- **Structure** based methods
  - *Identify substructures that are rare structurally*
  - Connectivity, attributes
- **Community** based methods
  - *Aim to identify community outliers (nodes)*
  - CODA …an **unsupervised learning algorithm**
  - GOutRank …searches for **a subset of relevant attributes**
- **Relational learning** based methods
  - *Exploit the relationships between the objects to assign them into classes*
  - **Naive Bayes models** for local attributes
  - **Probabilistic relational models** (PRMs)

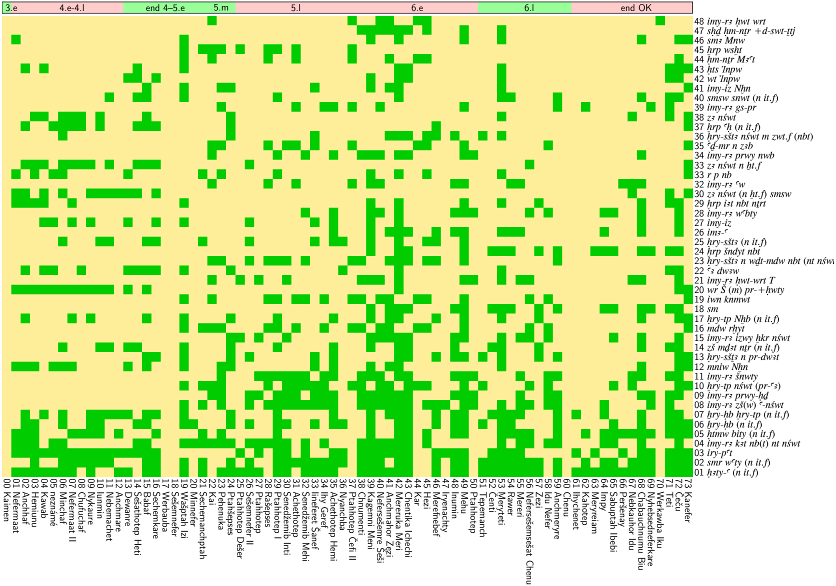
Anomaly Detection in Dynamics Graphs [ATK15]

- **Feature** based events
  - *"graph footprints" and metrics,*
  - Maximum Common Subgraph (MCS)
  - Graph Edit Distance (GED)
  - Hamming distance
- **Decomposition** based events
  - *matrix or tensor decomposition of the time-evolving graphs*
  - **Singular Value Decomposition (SVD)**
  - **Non-negative Matrix Factorization (NMF)**
  - Compact Matrix Decomposition (CMD)
  - Streaming Tensor Analysis (STA)
- **Community/cluster** based events
  - *graph communities over time*
  - **clustering, community detection, co-clustering**
  - **MDL-based, Bayesian anomaly detection method**
- **Window** based events
  - *"moving window analysis"*
  - k-step neighborhood

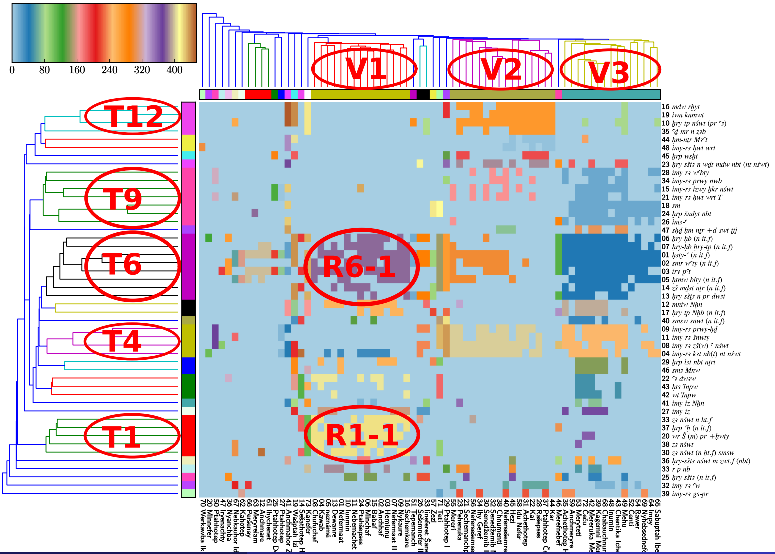
Rebellion [Eps02, Wil04, Wil99]

- This project models the rebellion of a subjugated population against a central authority.
- If the level of population wanders grievance against the central authority is high enough, and their perception of the risks involved is low enough, they openly rebel.
- The cops wander around randomly and arrest people who are actively rebelling.
- **Punctuated equilibrium** — periods of quiescence followed by periods of rebellion.

Titles of Viziers [DMBC17]

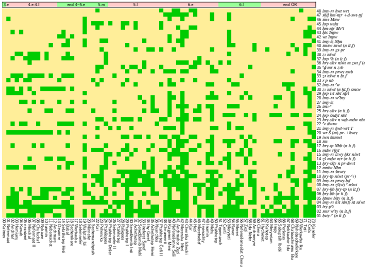


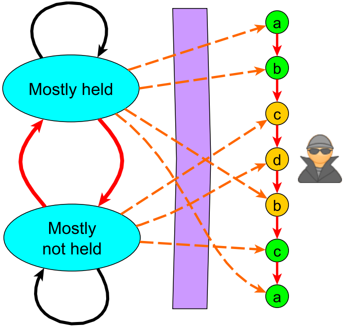
Titles of Viziers - Jaccard, Single Linkage Clustering [DMBC17, JD88]



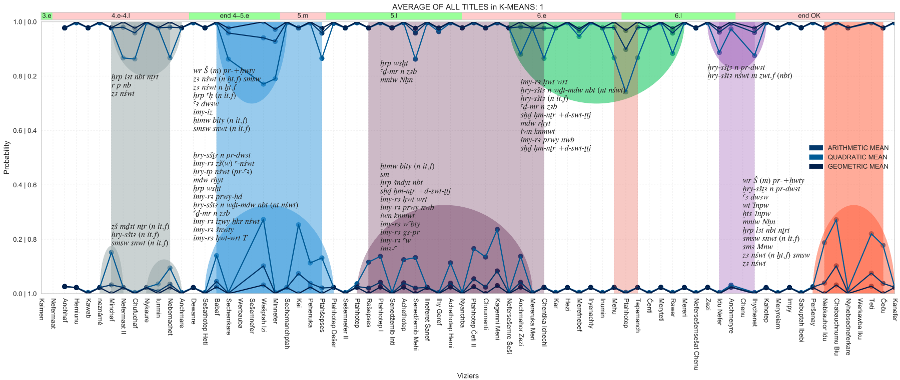
Hidden Markov Model on Titles [DMBC17]

- A sequence of viziers
- A sequence of appearances for each title
- A general model of title life (2 states)
- Focus on title occurrence change (red)
- A model of a title subset change
- Identification of periods when with a higher probability
  - A subset of titles started to appear
  - A subset of titles stopped to appear
- Identification of titles contributing to changes



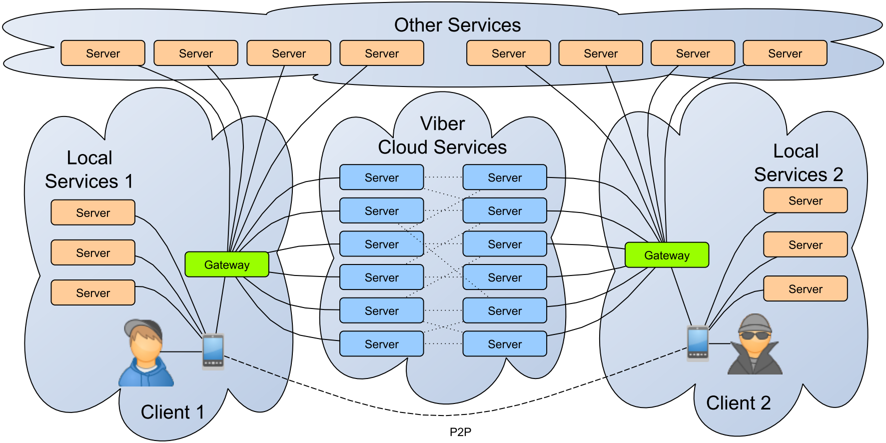


The Old Kingdom Administration Rise

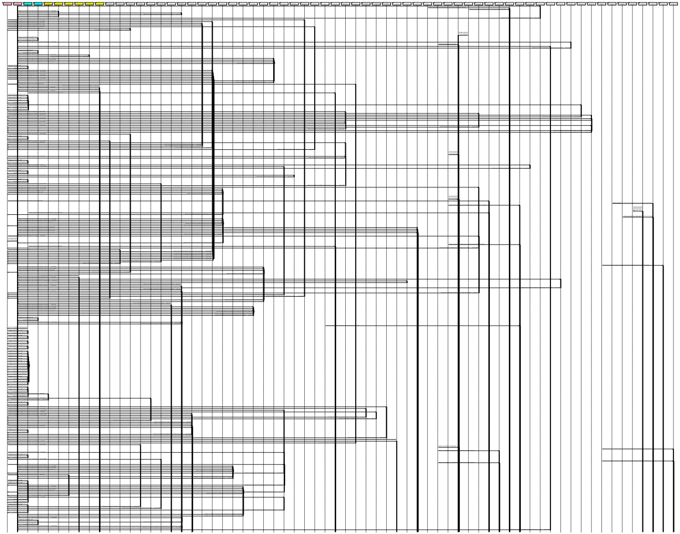


- Probability means of join changes in title occurrences
- Non-informative titles threshold (*feature selection*)
- Top – vanishing titles
- Bottom – rising titles
- Epochs match conclusions of Egyptologists

Exemplar (Viber) Environment [MBKK15]



Example Capture Characteristics - Message Sequences [MBKK15]

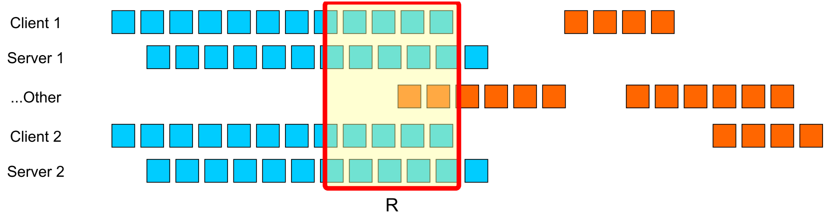


- 138882 PCAP blocks
- 1788 transport sessions
- 2 clients
- 22 viber.com servers
- 150 peers of 2 clients
- 5660 possible concurrent sessions
- How to analyze?

Concurrent Communication Detection [MBKK15]

Selection of IP nodes

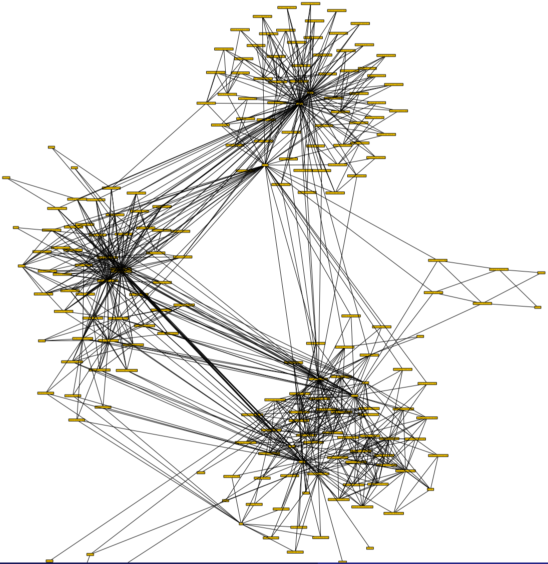
- *viber.com* servers $\rightarrow$ viber clients $\rightarrow$ other Viber servers
- Classified based on entropy based characteristics of TCP/IP distributions



$$s(a,b) = \frac{\sum_{\forall i,j: t_a[i]-t_b[j]<R} R/(t_a[i]-t_b[j])}{\sum_{\forall i,j: t_a[i]-t_b[j]<R} 1}$$

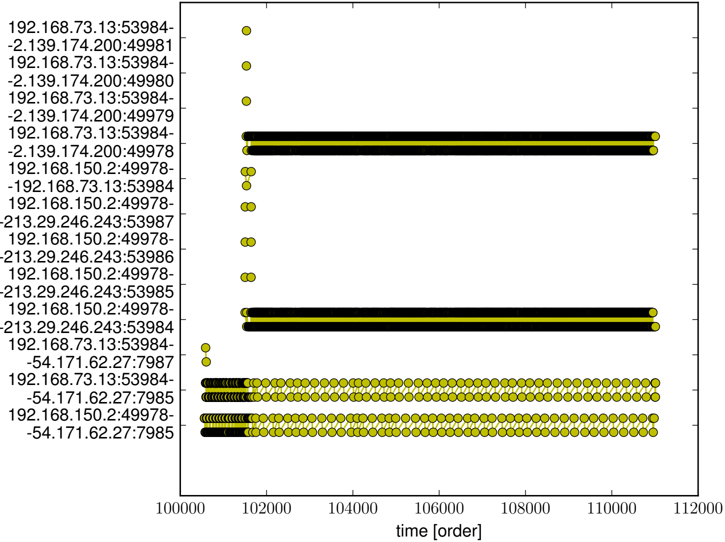
In our experiments: $R = 50ms, \quad s(a,b) > 0.001$

UDP Packet Sequence Concurrency as a Complex Network [MBKK15]



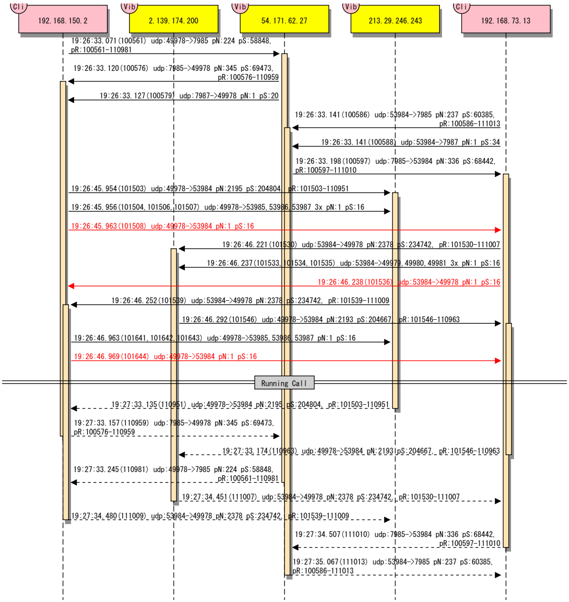
- Captures with two clients
- **Communities** of concurrent sessions
- Some clusters related to only one client
- Interesting clusters consist of nodes of **both** clients

UDP Packet Sequence Concurrency - Packet Timing [MBKK15]



- Signals
- Calls
- Keep-alive packets
- Direct client to client packets

Voice Call [MBKK15]



Conclusions

- A brief introduction to **anomaly and outlier detection**,
- **Complex networks** introduced as a representation of data with
  - Large-range dependencies,
  - Specific types of dependency topologies,
  - Large/huge volume of data.
- Anomalies might be detected using traditional **machine learning** methods with
  - Adjustments to huge data,
  - Special null models,
  - Special graph/network topological features
  - Community detection (topology/relation based $\times$ clustering)
- Applicable to many diverse domains

THANK YOU

References I

[Agg17] Charu C. Aggarwal. *Outlier Analysis*. Springer, second edition, 2017.

[ATK15] Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29(3):626–688, 2015.

[BAJ99] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1):173 – 187, 1999.

[BGLL08] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[BL94] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley & Sons, Kluwer Academic Publishers, Boston/Dordrecht/London, 1994.

[Bri95] Matt Britt. Partial map of the internet 1995, accessed 28.1.2014. `http://en.wikipedia.org/wiki/Wikipedia:Featured_picture_candidates/Internet_Map`, 1995.

[CBK09] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009.

[Col01] Stuart Coles. *An introduction to statistical modemodel of extreme values*. Springer, 2001.

[CRTV07] L. D. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56:167–242, January 2007.

[DM15] Veronika Dulíková and Radek Mařík. Social network analysis in the Old Kingdom society: the nepotism case. In *Presented at conference Abusir 2015, Czech Institute of Egyptology, Charles University, Prague, CZ*, 2015.

[DMBC17] Veronika Dulíková, Radek Mařík, Miroslav Barta, and Matej Cibuľa. HMM model vývoje a trendů správy země v období Staré říše. In *16. ročník konference Počítačová podpora v archeologii, Písek CZ, 29. - 31. května 2017*. Katedra archeologie Západočeské univerzity v Plzni, CZ, 2017.

References II

[EK10] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets. Reasoning About a Highly Connected World*. Cambridge University Press, July 2010.

[Eps02] Joshua M. Epstein. Modeling civil violence: An agent-based computational approach. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7243–7250, 2002.

[FMS] FMS. Social network analysis (SNA) diagram, al qaeda terrorist network, accessed 28.1.2014.

[GDZ+15] Dehua Gao, Xiuquan Deng, Qiuhong Zhao, Hong Zhou, and Bing Bai. Multi-agent based simulation of organizational routines on complex networks. *Journal of Artificial Societies and Social Simulation*, 18(3):17, 2015.

[Gru50] Frank E. Grubbs. Sample criteria for testing outlying observations. *Ann. Math. Statist.*, 21(1):27–58, 03 1950.

[Gru69] Frank E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.

[Gru05] Peter Grunwald. *Advances in Minimum Description Length: Theory and Applications*, chapter A tutorial introduction to the mimimum description principle, pages 3–79. Cambridge, MA, MIT Press, 2005.

[HA04] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[Haw80] D. Hawkins. *Identification of Outliers.* Chapman and Hall, 1980.

[HK13] Stephen J. Hardiman and Liran Katzir. Estimating clustering coefficients and size of social networks via random walk. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 539–550, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[JD88] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[Jen55] A.F. Jenkinson. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of Royal Meteorological Society*, 81:58–171, 1955.

References III

[KKZ10] Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Outlier detection techniques tutorial at 16th acm sigkdd conference on knowledge discovery and data mining (kdd 2010), washington, dc, 2010., 2010. Accessed 18 June 2017.

[Mar17a] Radek Marik. *Efficient Genealogical Graph Layout*, pages 567–578. Springer International Publishing, Cham, 2017.

[Mar17b] Radek Marik. Threshold selection based on extreme value theory. In *accepted for Mechatronics 2017*, 2017.

[MBKK15] Radek Mařík, Pavel Bezpalec, Jan Kučerák, and Lukáš Kencl. Revealing viber communication patterns to assess protocol vulnerability. In *2015 International Conference on Computing and Network Communications (CoCoNet). Leonia, NJ 07605: EDAS Conference Services*, volume ISBN 978-1-4673-7308-1, pages 502–510, 2015.

[MD15] Radek Mařík and Veronika Dulíková. *Mathematical Formalization of Society Complexity*, chapter Povaha změny: Bezpečnost, rizika a stav dnešní civilizace, pages 98–129. Praha Vyšehrad, ISBN 978-80-7429-641-3, 2015. (in Czech).

[New06] M E Newman. Modularity and community structure in networks. *Proc Natl Acad Sci U S A*, 103(23):8577–8582, June 2006.

[New10] M. Newman. *Networks: an introduction*. Oxford University Press, Inc., 2010.

[PASP09] Jose M Peregrin-Alvarez, Chris Sanford, and John Parkinson. The conservation and evolutionary modularity of metabolism. *Genome Biology*, 10(6), June 2009.

[Ris78] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, September 1978.

[Roc12] Claudio Rocchini. Centrality. `http://en.wikipedia.org/wiki/File:Centrality.svg`, November 2012.

[RP13] James Risen and Laura Poitras. N.S.A. gathers data on social connections of U.S. citizens, September 2013.

[STE07] Jerry Scripps, Pang-Ning Tan, and Abdol-Hossein Esfahanian. Node roles and community structure in networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 26–35, New York, NY, USA, 2007. ACM.

References IV

[SWJR07] Xiuyao Song, Mingxi Wu, Christopher Jermaine, and Sanjay Ranka. Conditional anomaly detection. *IEEE Trans. on Knowl. and Data Eng.*, 19(5):631–645, May 2007.

[Weh13] Stefan Wehrli. Social network analysis, lecture notes, December 2013.

[Wil99] U. Wilensky. Netlogo, 1999.

[Wil04] U. Wilensky. Netlogo rebellion model, 2004.

[WS98] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of /'small-world/' networks. *Nature*, (393):440–442, June 1998.