Winter School on Data Analytics (DA-2016), Nizhny Novgorod , Russia

Big data: towards building custom in-house solutions Applications of techno-social systems in economy and governance

Alexander Troussov, Ph.D.

Mathematical Methods for Social Network Mining Laboratory,
The Russian Presidential Academy
of National Economy and Public Administration.

AGENDA

- WHAT
  - 3 projects in a nutshell, QA posed by discussion on these projects

- QUESTIONS: What is knowledge, network models of techno-social systems, network mining and algorithms of graphs…
- WHY – why network models, etc..
- ANSWERS in a nutshell:
  Network models are good for nuanced empirical environment
  "Finite-difference" method on networks makes sense
- HOW
  - Details of some projects are below
    this is not only a theory, is not necessarily the best approach, but it works

Collaboration

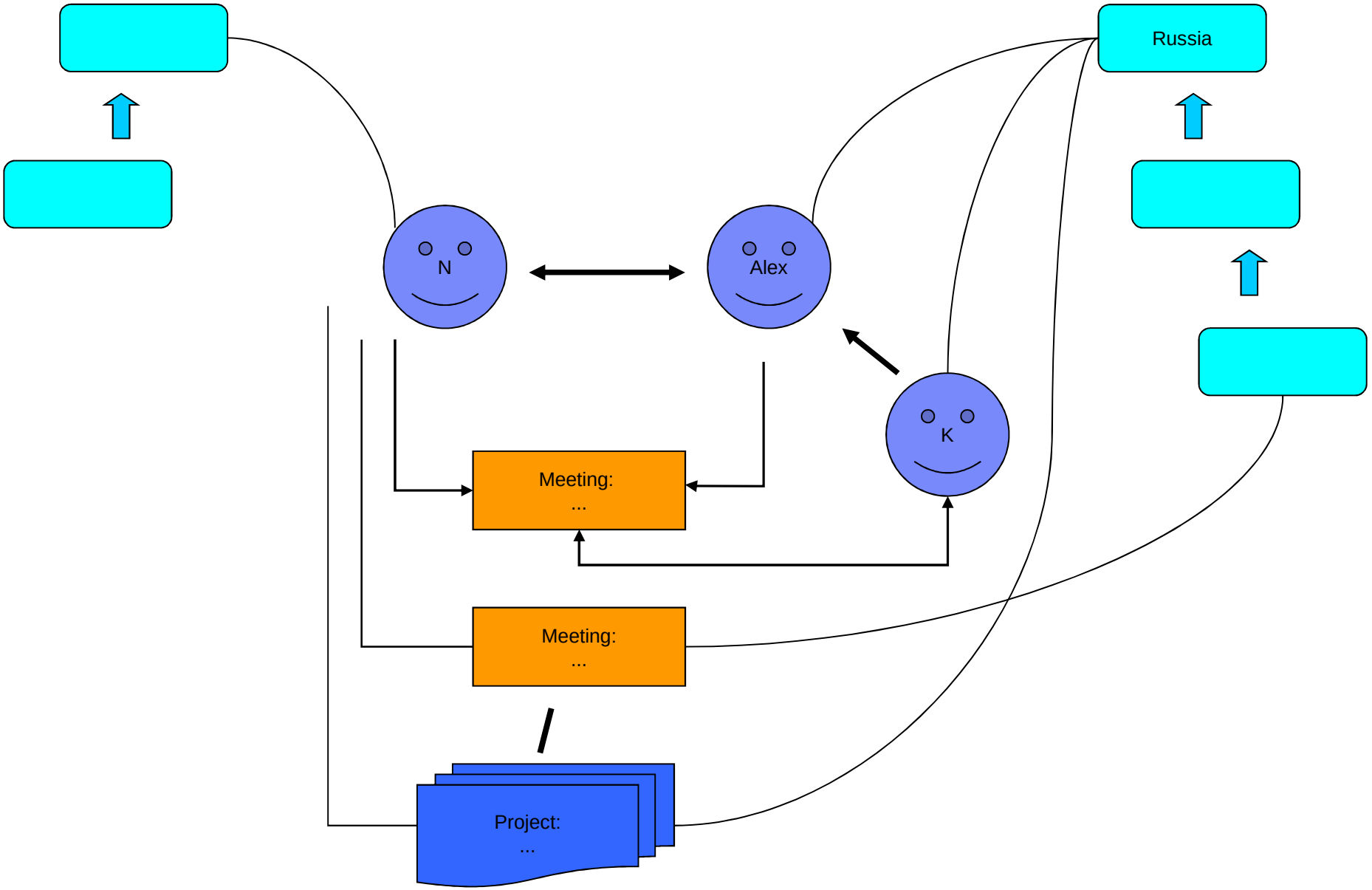
- I'm working on an edited book (related to the topics of this Winter School)
- Should you be interested
  - In collaboration with my Lab (we are hiring)
  - To work with me on curricula for teaching BIG DATA, big data analytics
  - To advice me regarding suitable topics for the book
  - To receive calls for chapters
  - To write chapters and get Thomson Reuters, Scopus etc publications
- Please contact me at

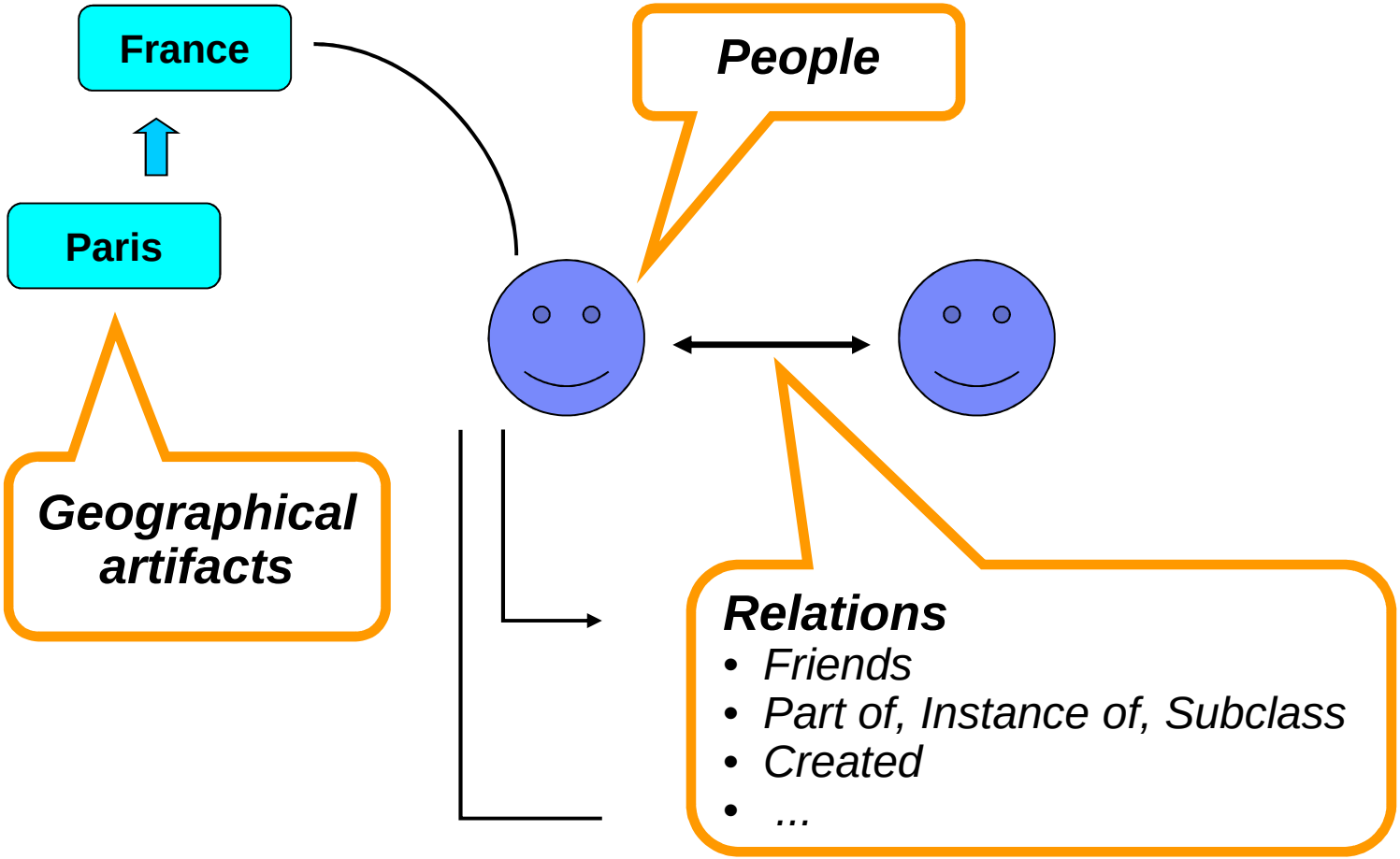
troussov@gmail.com
Alexander Troussov

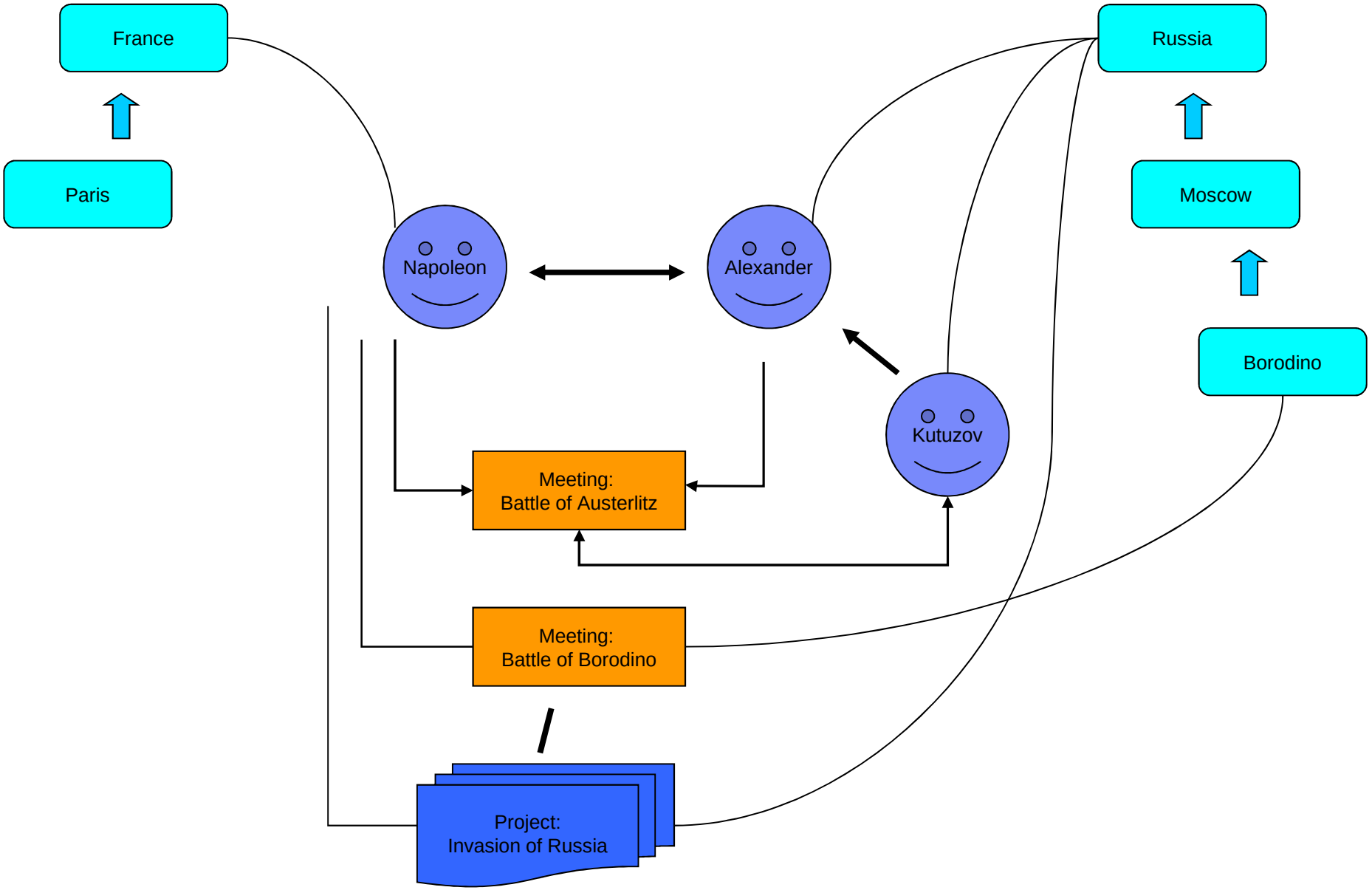
Three of my recent projects

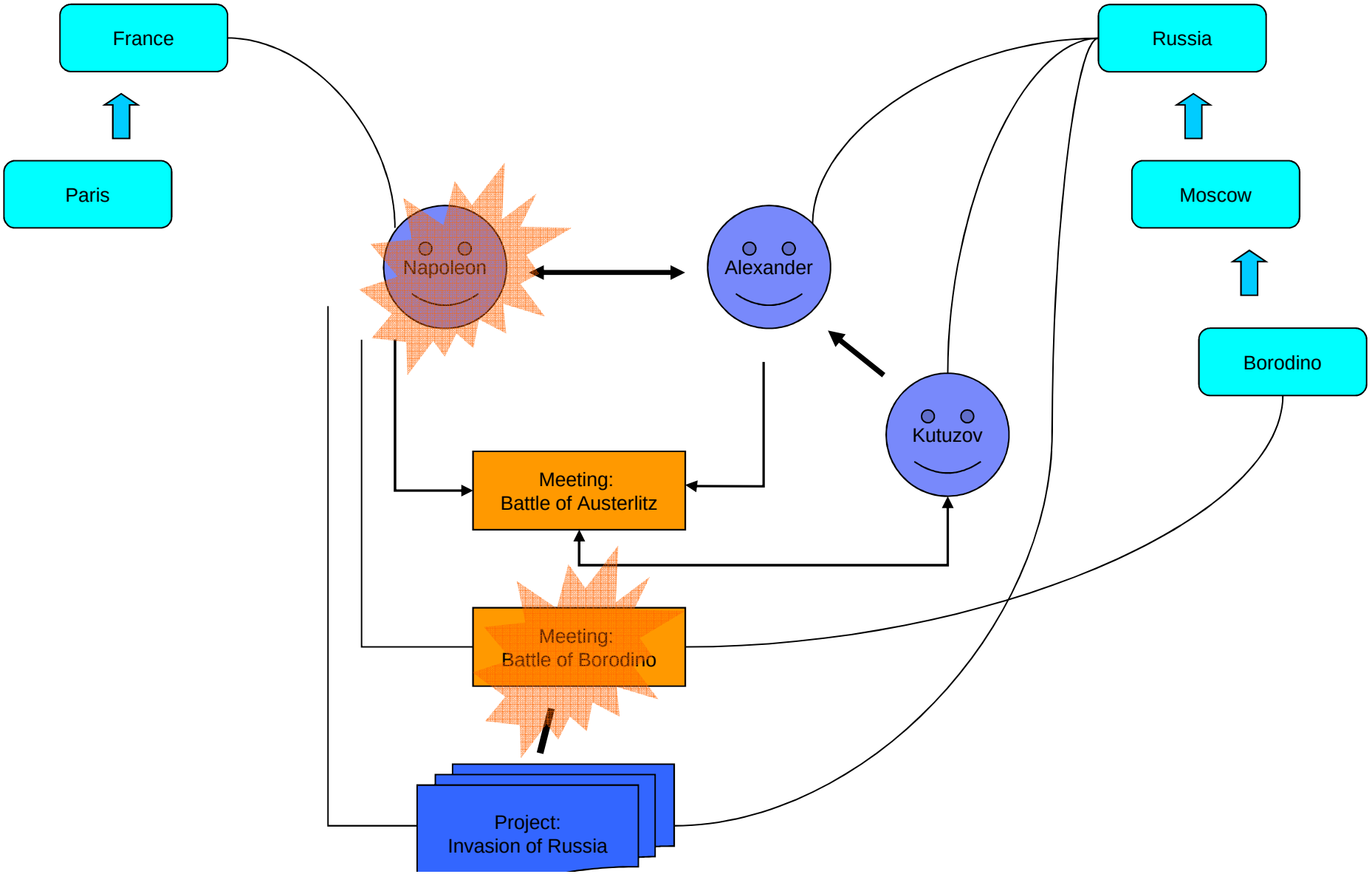
What does the following diagram represent?

- Data from Facebook?
- An ontology?
- Collocation of terms in a text?
- …


Russia
N
Alex
K
Meeting:
...
Meeting:
...
Project:
...


France
Paris
Geographical artifacts
People
Relations
• Friends
• Part of, Instance of, Subclass
• Created
• ...


France
Paris
Napoleon
Alexander
Russia
Moscow
Borodino
Kutuzov
Meeting:
Battle of Austerlitz
Meeting:
Battle of Borodino
Project:
Invasion of Russia


France
Paris
Napoleon
Alexander
Russia
Moscow
Borodino
Kutuzov
Meeting:
Battle of Austerlitz
Meeting:
Battle of Borodino
Project:
Invasion of Russia

What does the following diagram represent?

- Whatever it represents, it can be easily used,
  for instance, for recommendations like whom Napoleon should invite to this next business meeting “Battle of Borodino”, related to his new project “Invasion of Russia”, as is shown on the previous slide
- To compute recommendation
  - Put the initial activation at the point of interests
    - What - The meeting
    - For whom this recommendation – for Napoleon
  - Propagate this activation
  - Check Constraints on the recommendations – People
  - Select most activated nodes corresponding to People
  - Other highly activated nodes – use for explanations of the recommendation
- Computed recommendation implicitly takes into account most of the requirements for good recommendations for this invitation
  - Somebody who is an expert in Russia
  - Who is likely to accept the invitation
- All the information for such recommendations is encoded in the topology of the network data model

LESSONS LEARNED FROM THIS USE CASE

Analysis

- This work has been done in IBM
  - Enterprise collaboration
  - Customer engagements
  - 17M Euro Integrated EU Project
- Invites considerations on
  - What are Techno-Social Systems
  - Why network models
  - What is the specific of these networks
  - What kind of graph methods needed and how to generalize the zoo of applicable graph algorithms…

Answers

- Partial answers were given in my recent publications
- Are presented on next slides

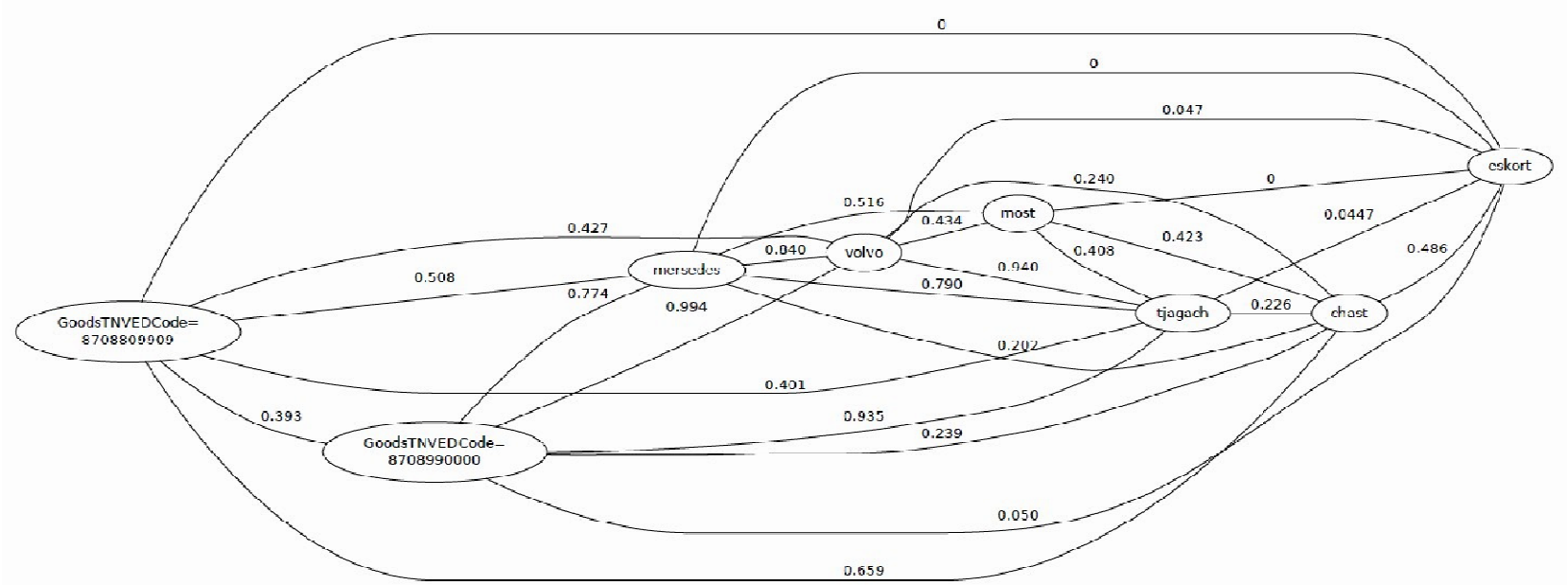
ANOTHER USE CASE

DATE MODELLING

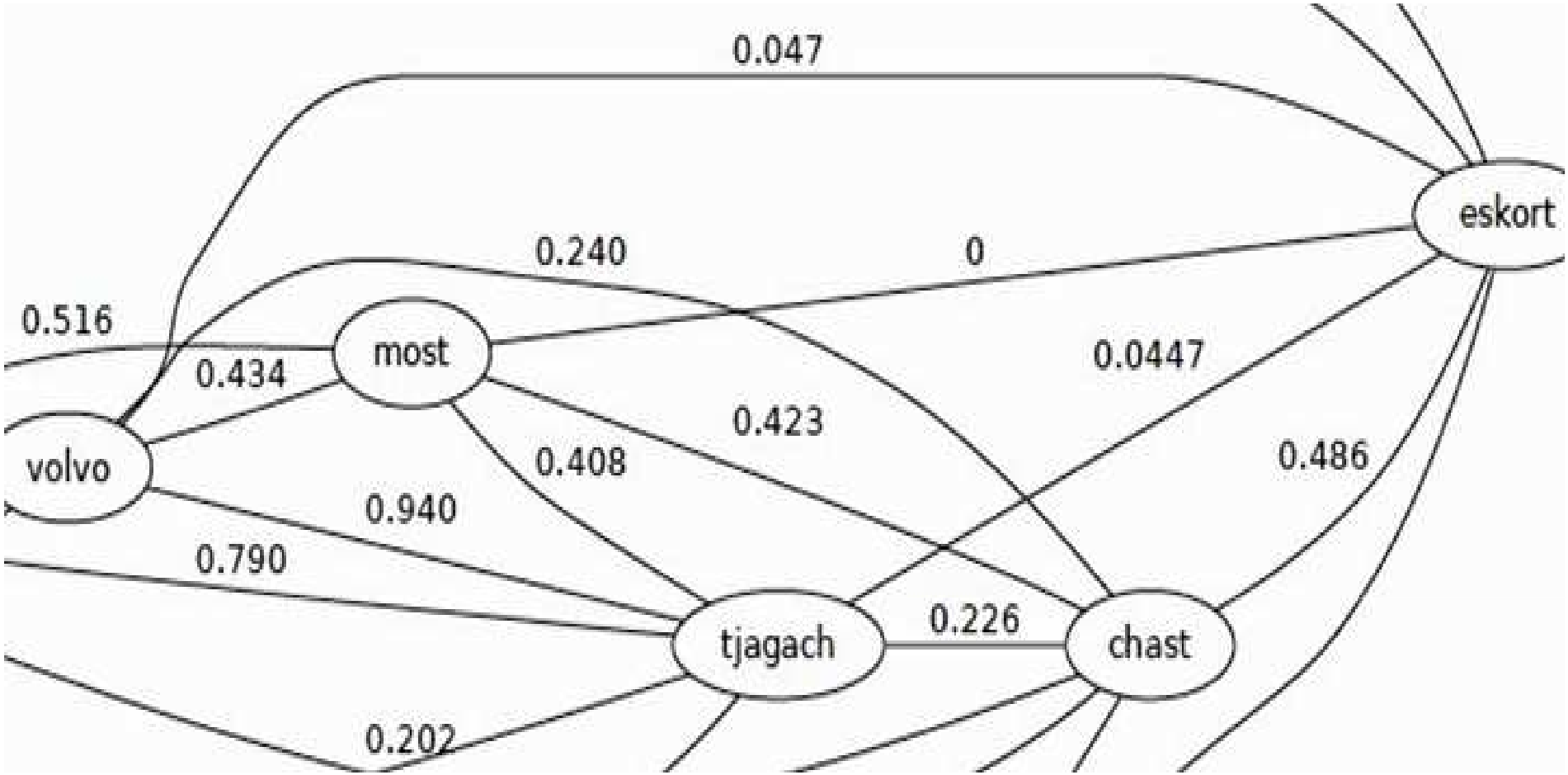
Modelling by multidimensional network

- *Multidimensional …*
  - *– having or involving or marked by several dimensions or aspects*
  - *–… A multi-dimensional database is structured by a combination of data from various sources*
- Multidimensional Network - nodes represent
  - –various abstract codes, for instance, assignment of the security escort, Customs Codes (Commodity codes, GoodsNTVEDCodes)
  - –numerical values measured in kg, seconds, USD etc,
  - –Words, natural language descriptions
  - –Actors (consignee, consignors, carriers)

Modelling by multidimensional network



- The fragment of the network which represents the data from customs declarations.
- Entities (Cells of the table shown before), after preprocessing are merged into one network, where nodes represent words, abstract codes (including assignment of the security escort)
- If two entities are met at least once in the same shipment document, the corresponding pair of nodes is connected by an arc. The weight of that arc represents how frequently the two entities are met in shipment documents (i.e the number of co-occurrences divided by the number of items)



- The fragment of the network which represents the data from custom declarations. The are two types of nodes on this figure: words and security escort tag.
- This network shows, for example, that the word “tjagach” (“тягач” in Russian; rooter, tractor in English) was met in good descriptions which require armed escort in 0.0447% cases. The shipments which has in the description the word “mersedes” (“мерседес” in Russian; Mercedes-Benz international) never required armed escort, while “volvo” required escort in 0,047% of cases. Shipments with parts, details of something (“chast”, “часть” in Russian), frequently were escorted.

Novel graph mining method

- We show that a wide range of graph-based algorithms popular in various application domains use the idea of propagation between nodes
- We discuss drawbacks and limitations of these algorithms belonging to the class of network flow algorithms (Borgatti 2005)
- To overcome these limitations, we are developing a much broader class of “physics-inspired” algorithms where the interaction between neighbor nodes could not always be interpreted as a flow producing a diffusion like process.
  - we show that iterative computational schemes similar to those used in network flow algorithms have been used for a long time in finite element analysis to solve physical problems and discuss a class of “physics-inspired” algorithms
- And show the road map to combine physics-inspired algorithms with “logic-inspired” algorithms on graph based on the extension of cellular automata procedure that determines the new state of a cell for the next generation
- We show the applications of these novel class of algorithms to core tasks of network mining – centrality measurements and clustering

PROPAGATION ALGORITHMS AND THEIR USE

- Formally, solution of many network data mining tasks boils down to the following problem: Given an initial function $F_0(v)$ on the network nodes, construct the function $F_{lim}(v)$ which provides the answer.
  - In different domains the function F_0 could be referred to as the initial conditions, the initial activation, semantic model of a text, etc.
  - In ontology based text processing, the initial function F_0 is the semantic model of a text w.r.t. to the knowledge: for instance, $F_0(v)=0$ if the concept v is not mentioned in the text, $F_0(v)=n$ if the concept v is mentioned n times.
    The function $F_{lim}(v)$ should show the foci of the text; for instance, $Argmax(F_{lim})$ is the most important focus of the text, while $F_{lim}(Argmax(F_{lim}))$ is the numerical value of the "relevancy"
  - In IR, the link analysis (such as Google's PageRank) ranks web pages based on the global topology of the network by computing $F_{lim}(v)$ using the iterative procedure where the initial condition is that all web pages are equally "important" ($F_0(v)\equiv 1$),

PROPAGATION ALGORITHMS AND THEIR USE (Cont.)

- Computationally efficient and scalable algorithms usually compute the function F_{lim} (which provides the "answer") using iterations: on each iteration the value of $F_{n+1}(v)$ is computed depending on the values of the function F_n on the nodes connected to the node v
  - Very broad range of algorithms including Google's PageRank, spreading activation, computation of eigenvector centrality using the adjacency matrix.
- Most of the mathematical algorithms behind such iterative computations are propagation algorithms (the "network flow" algorithms):
  they are based on the idea that something is flowing between the nodes across the links, and the structural prominence of nodes could be explained and computed in terms of incoming, outcoming and passing through traffic
- Similar iterative computational schemes have been used for long time in finite element analysis to solve physical problems including propagation of heat, of mechanical tensions, oscillations, etc.
  - Although finite element analysis automata usually perform on rectangular (cubic, etc.) grids, the extension to arbitrary networks is feasible.

PROPAGATION ALGORITHMS AND THEIR USE (Cont.)

- However, the interaction between the material points in mechanics could not always be described as a flow, and such interactions could model more complex processes than diffusion
  - For instance, one dimensional heat transfer equations can be numerically simulated on a one-dimensional mesh by iterations. On each iteration recomputation is based on the formula below:

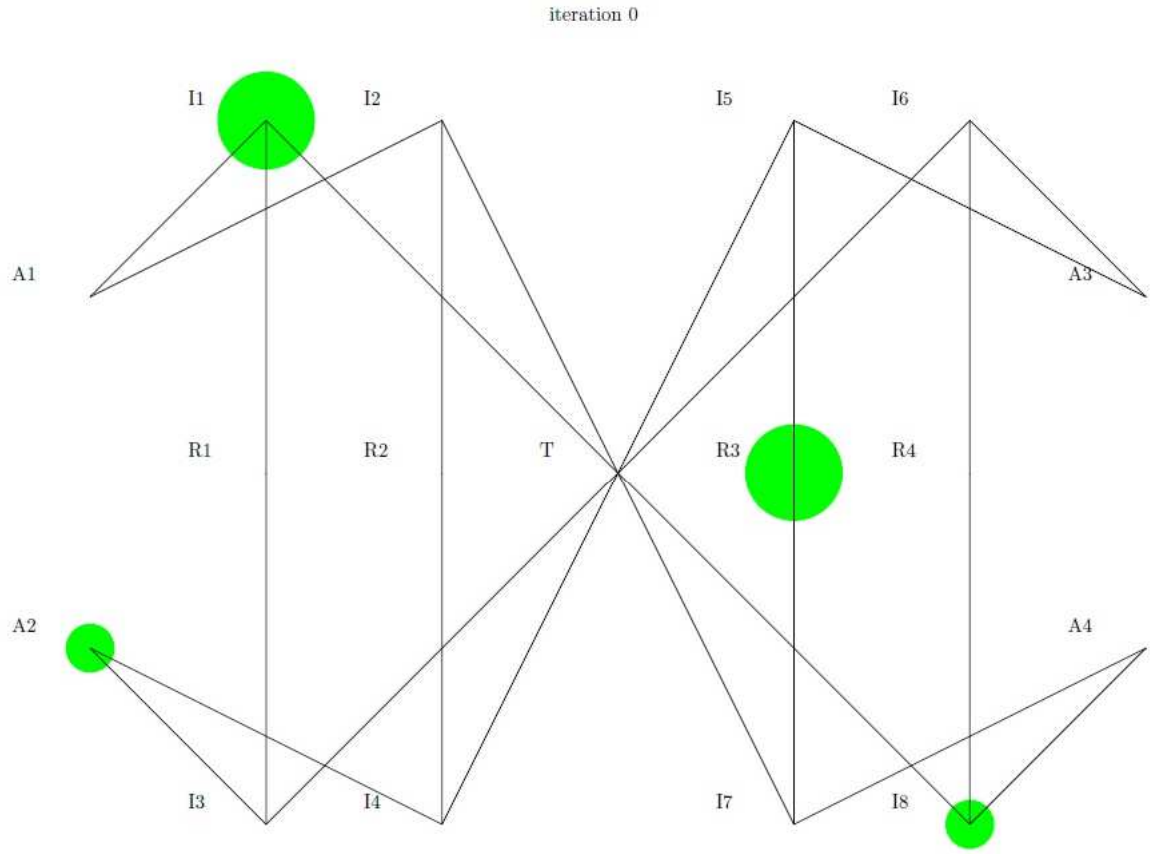
    $$F_{new}(v) = (F(RightNeighbour(v)) + F(LeftNeighbour(v))) / 2$$

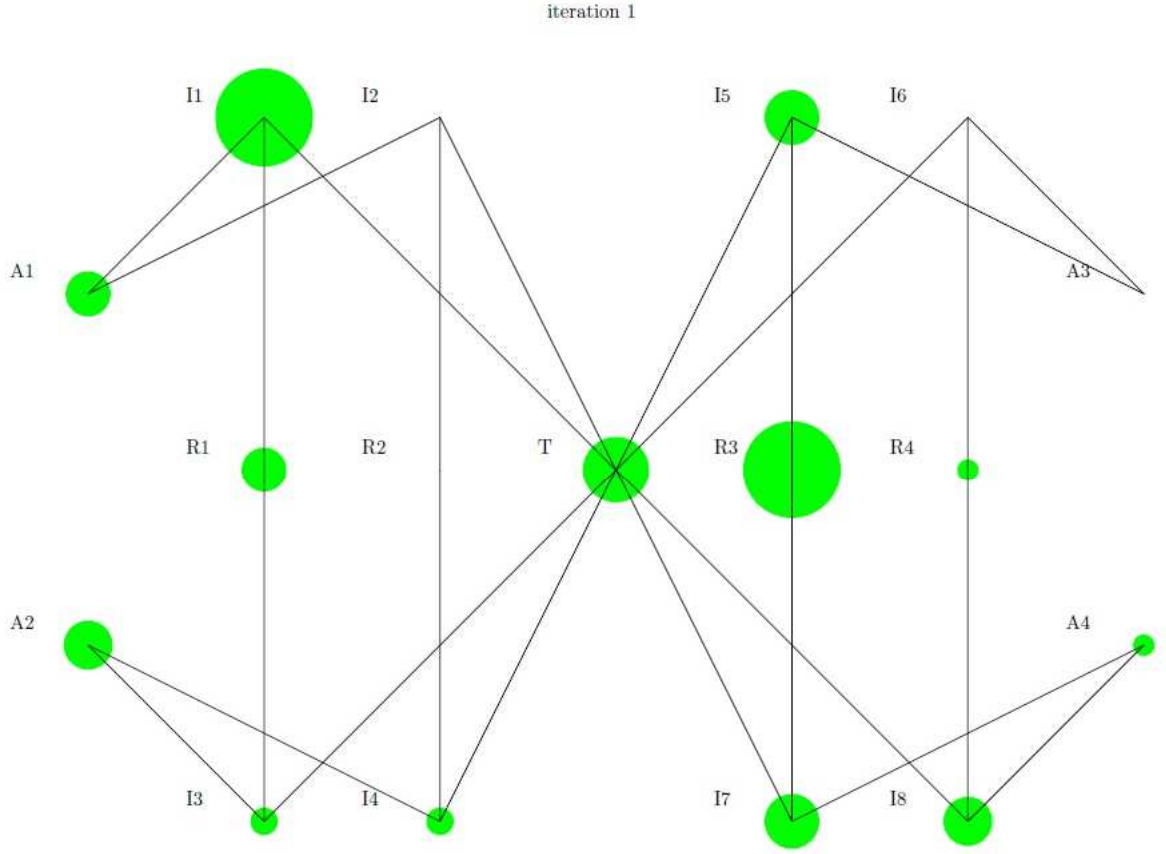
    This linear equation confirms the perception of the heat transfer as a flow: on each iteration the heat – the value of the function F – flows from nodes to the neighbour nodes.

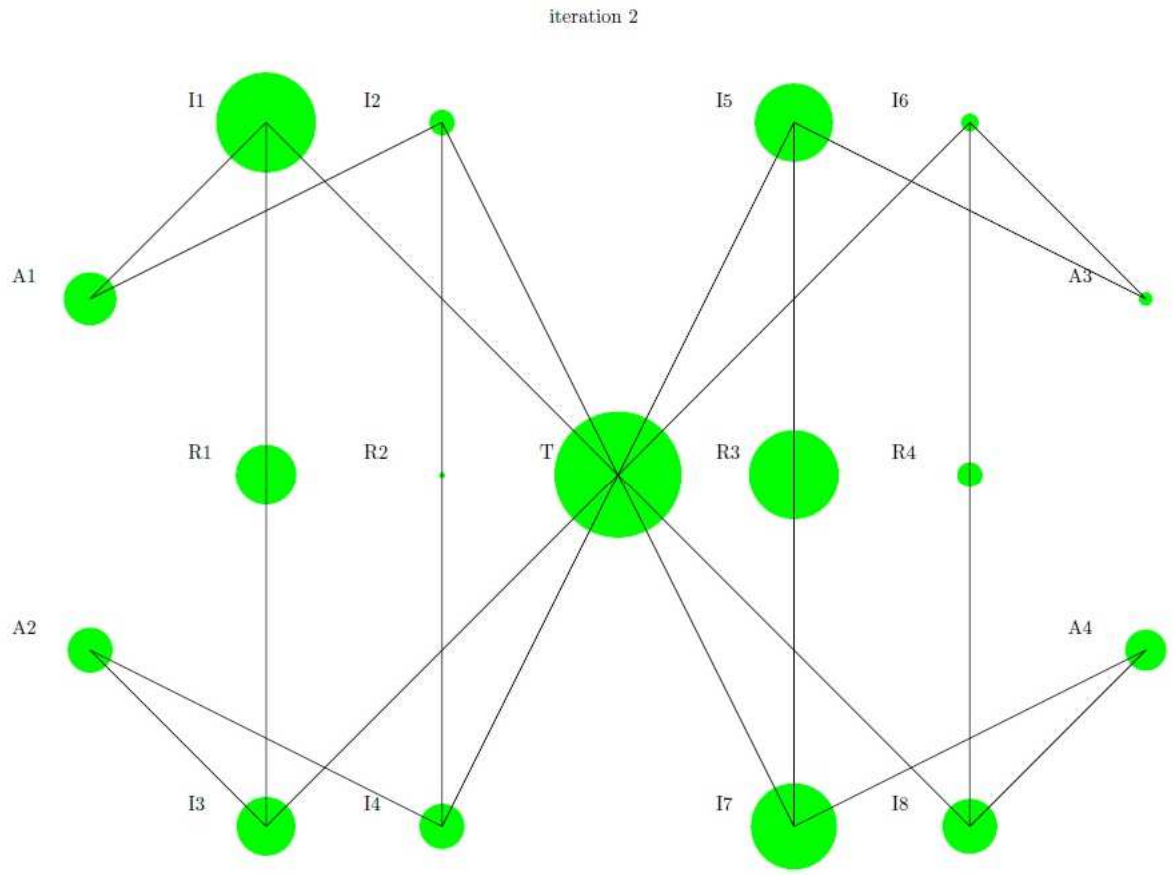
    In physics, a conservation law states that the amount of heat in an isolated physical system does not change as the system evolves, so “move mechanism” in heat propagation is a transfer. In network theory applications, network flow could be also done by “copy mechanism”, such as replication in spread of deceases.
  - At the same time, in physics, many processes can not be interpreted as a flow and can not be described by a function of one real variable. For instance, to simulate the behavior of an oscillating string one needs to operate with three values at each node - position, mass and velocity of the material point corresponding to the node. And none of these properties “flow” to the neighbors.

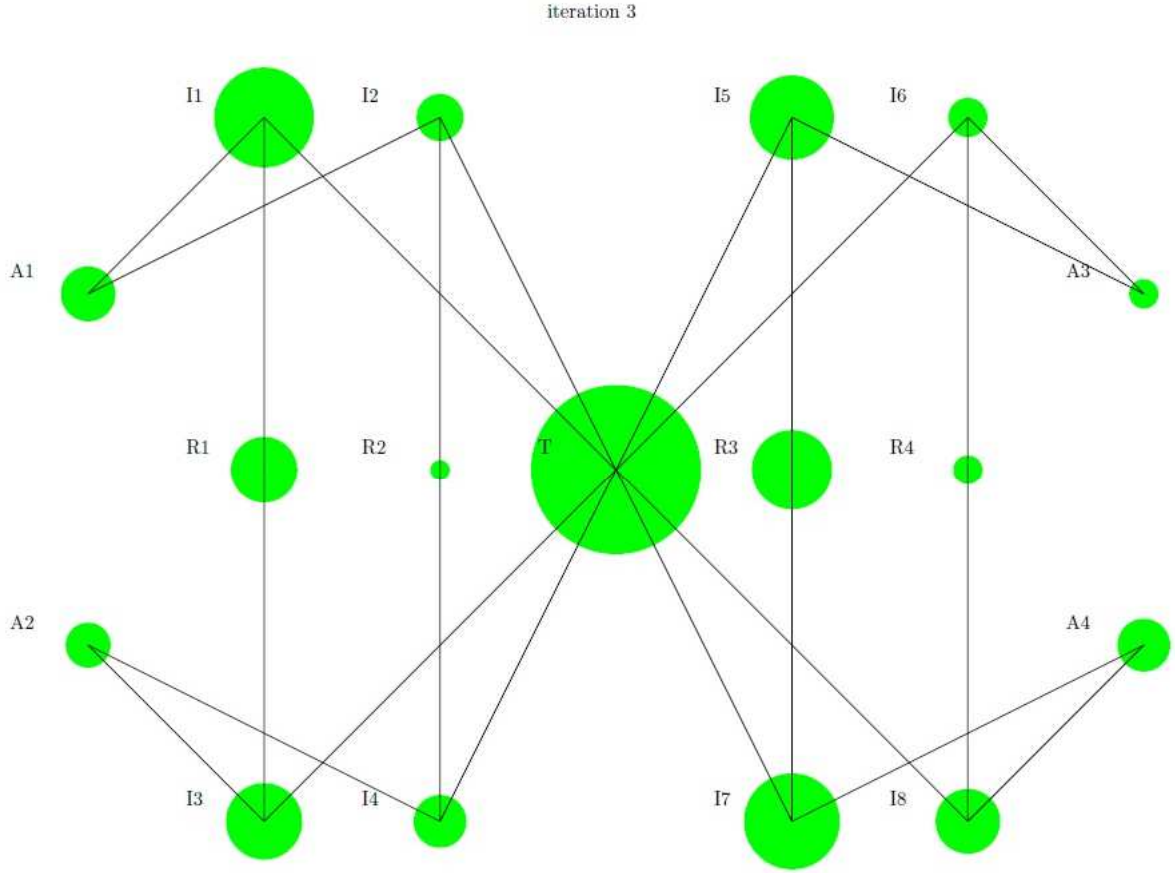
PROPAGATION ALGORITHMS AND THEIR USE (Cont.)

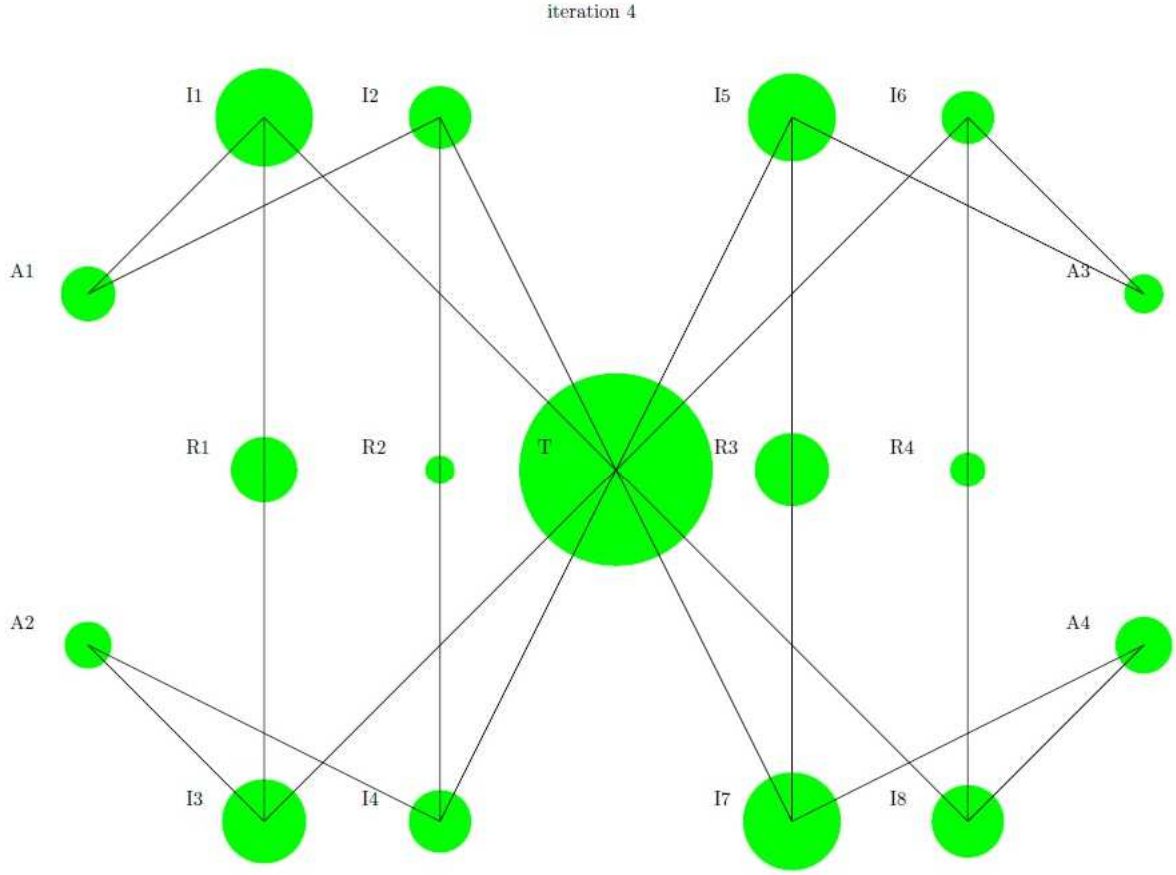
- Illustration how it works
  Graph mining methods are the same as in case of recommendations Whom to invite to the meeting discussed above

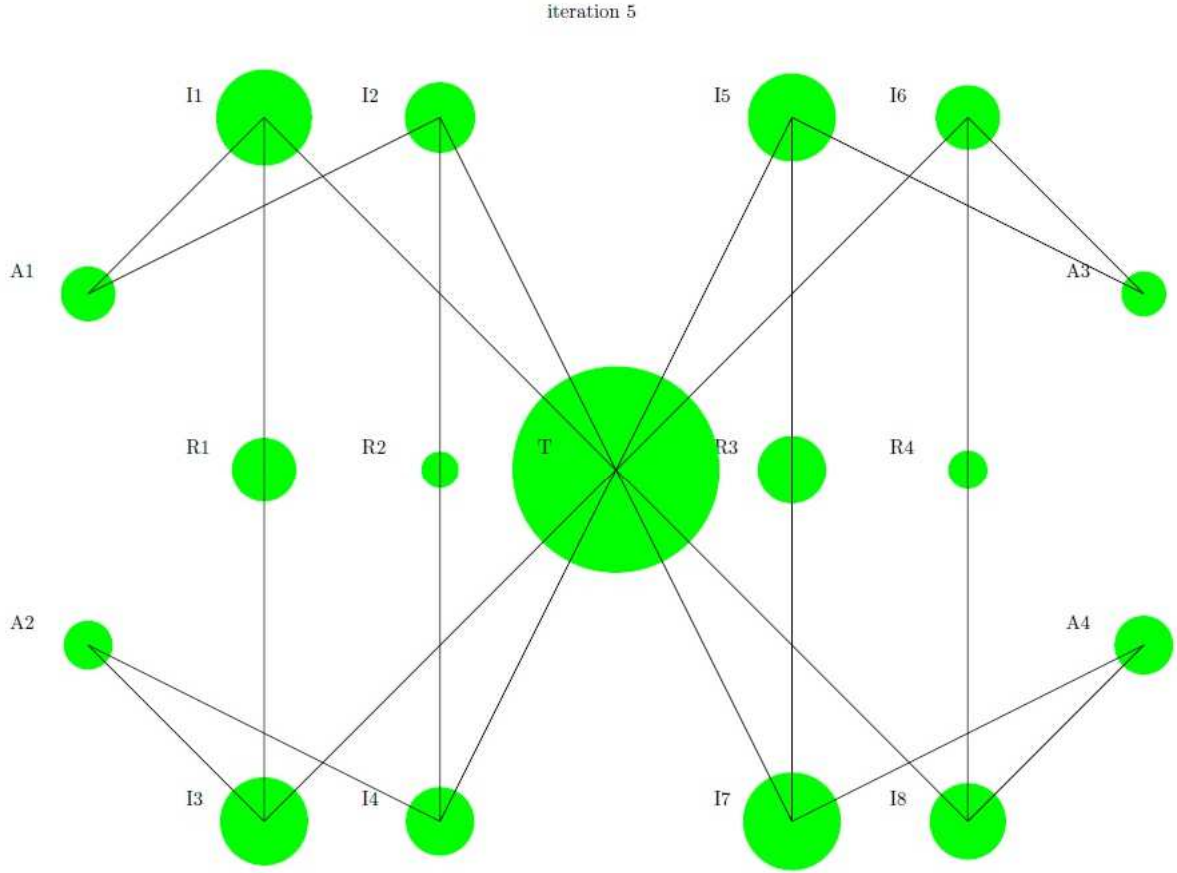

iteration 0
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
A4
I3
I4
I7
I8

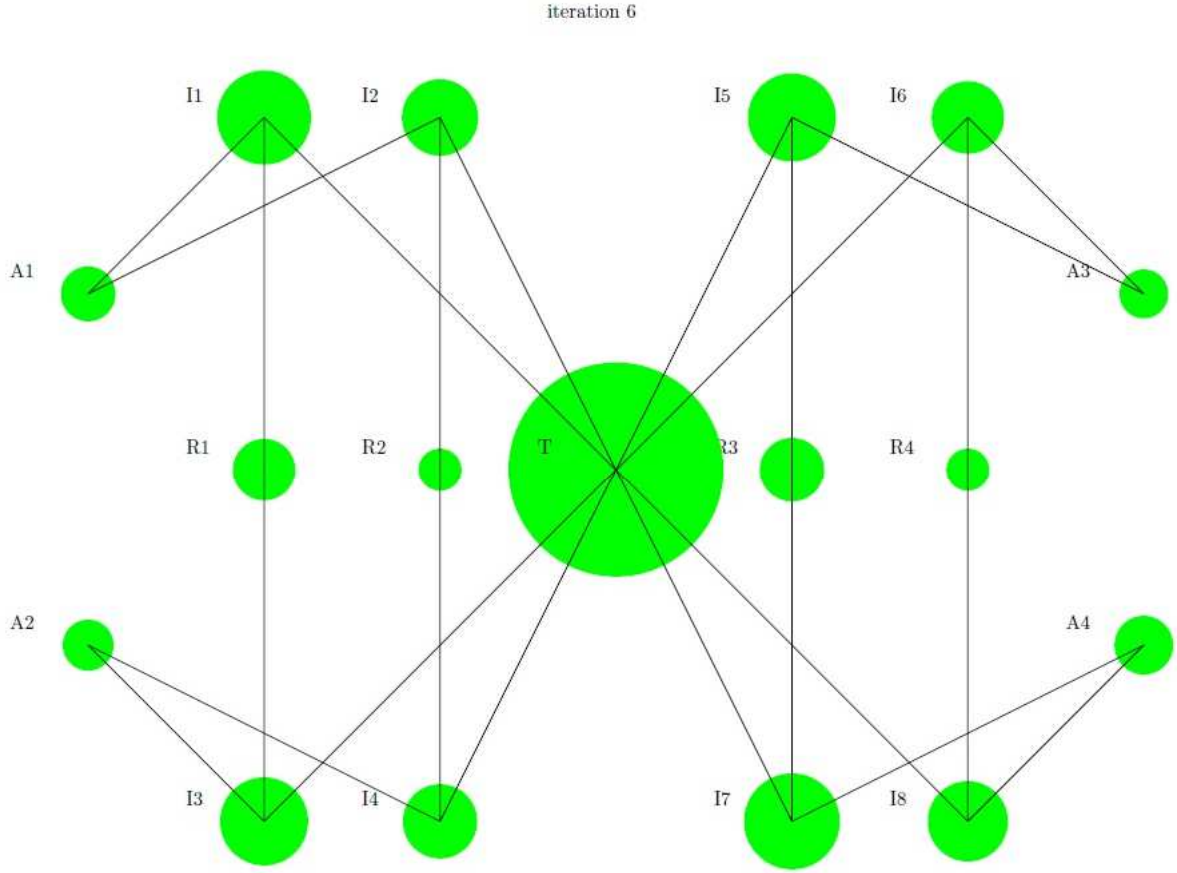

iteration 1
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
A4
I3
I4
I7
I8


iteration 2
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
A4
I3
I4
I7
I8


iteration 3
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
A4
I3
I4
I7
I8

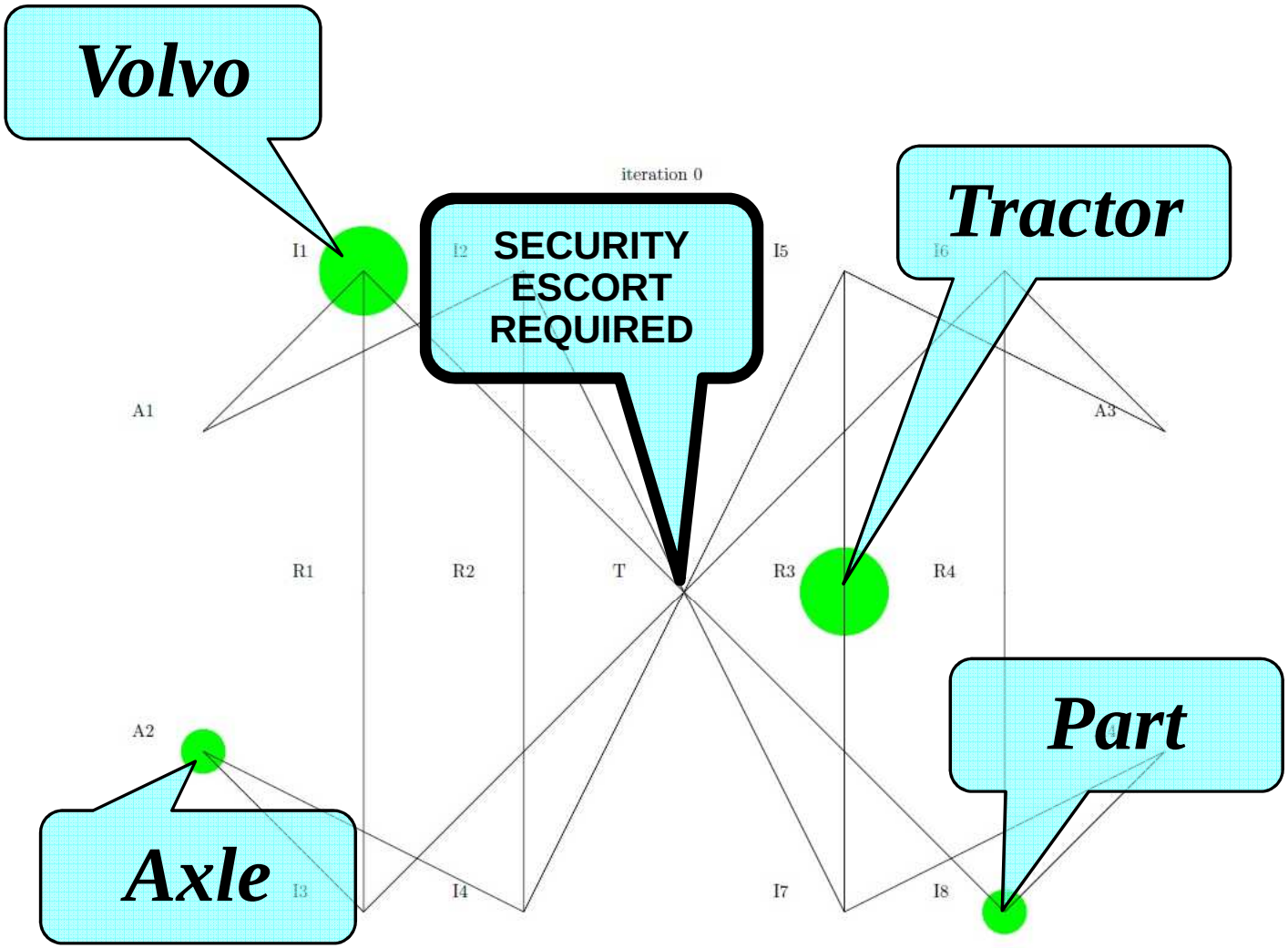

iteration 4
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
A4
I3
I4
I7
I8

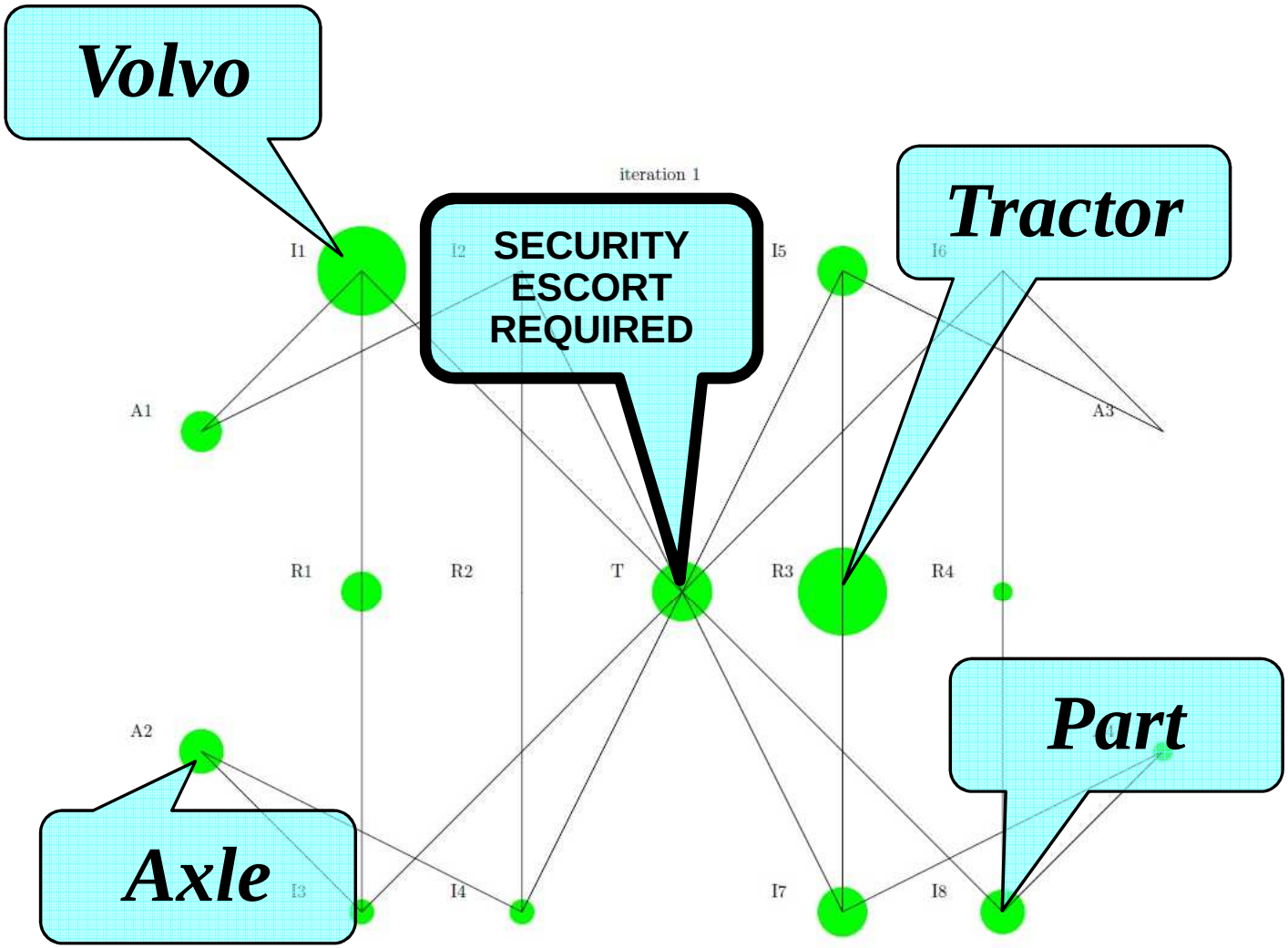

iteration 5
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
A4
I3
I4
I7
I8

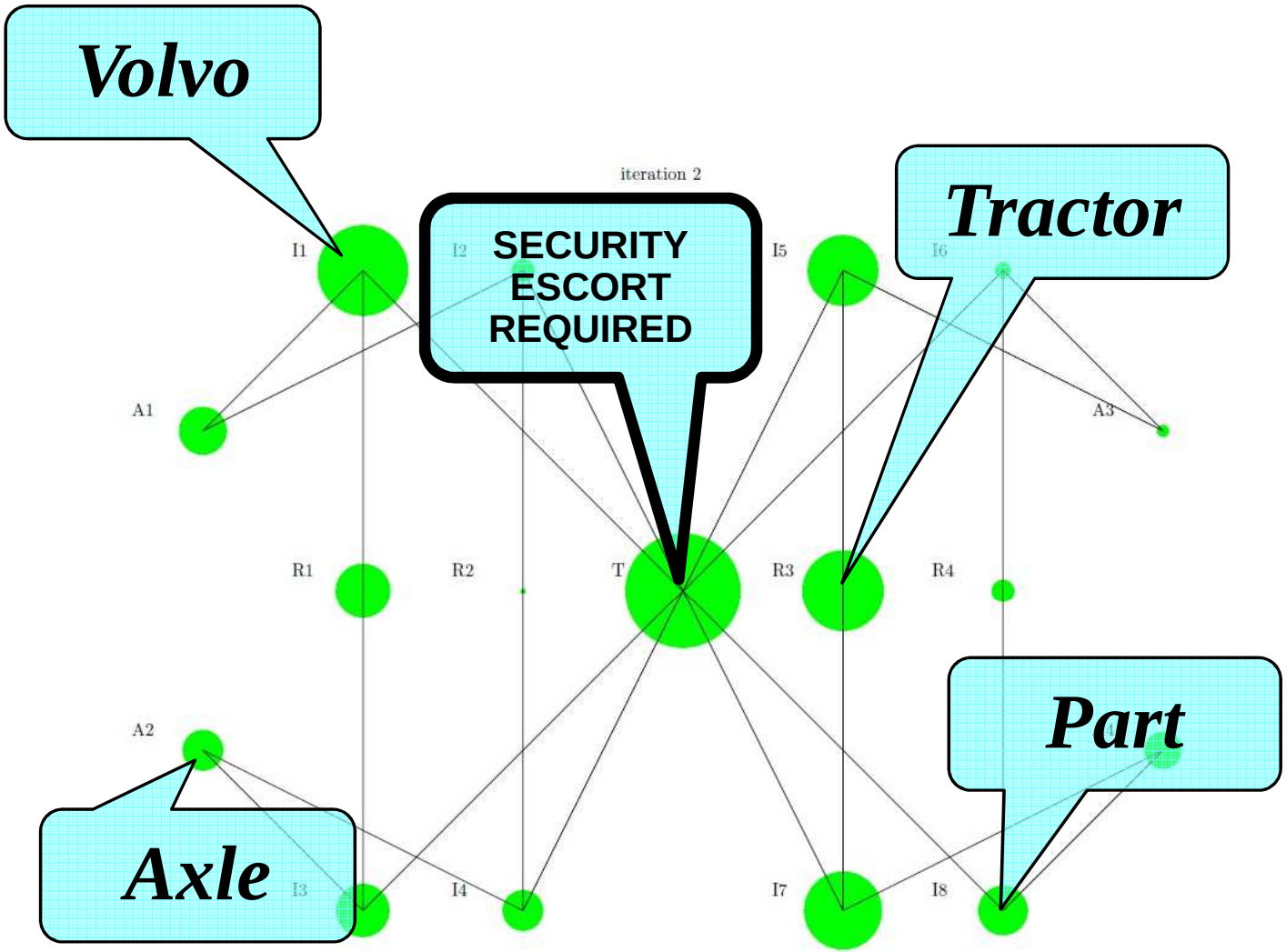

iteration 6
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
A4
I3
I4
I7
I8

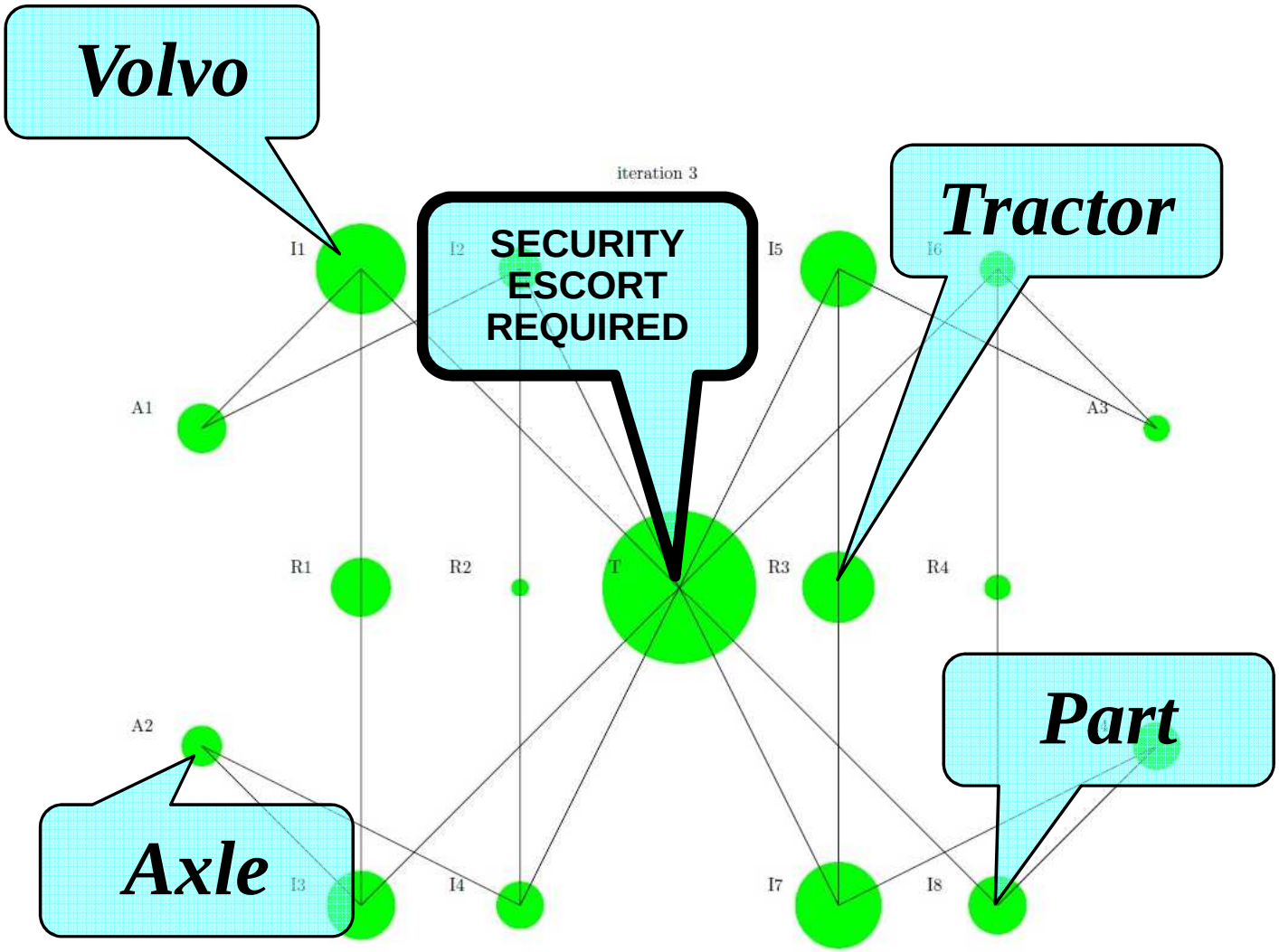
PROPAGATION ALGORITHMS AND THEIR USE (Cont.)

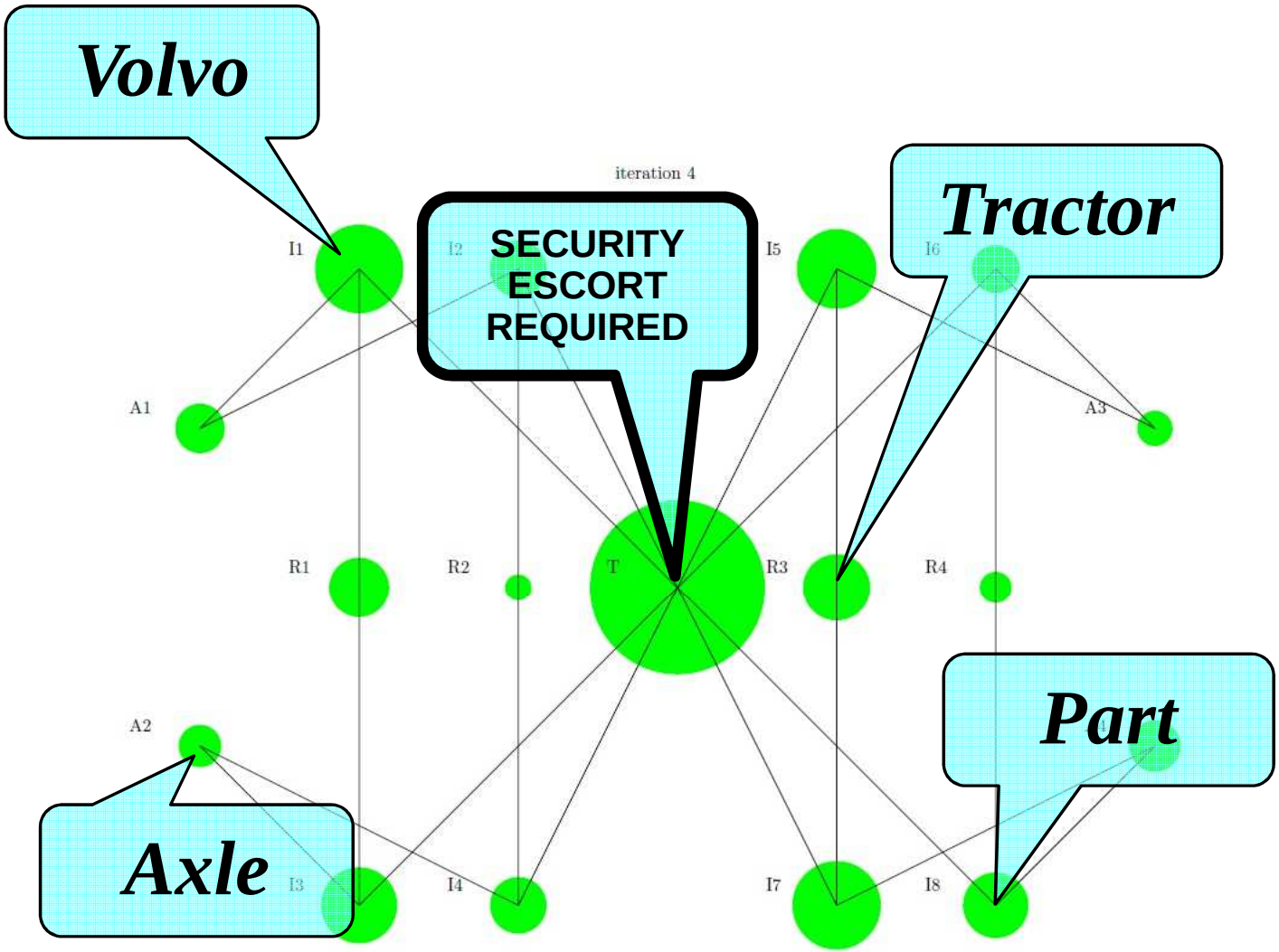
- Illustration how it works for recommendations related to Customs Declarations
  See next slide
- The data are converted to a particular network model
- A new customs declaration is mapped into this model
  thus providing initial distribution of the activation on the network
  The nodes marked by
  VOLVO, TRACTOR, AXLE, PART
  in our case
- The activation is propagated
- And measured at some nodes of interests
  SECURITY ESCORT REQUIRED in this case

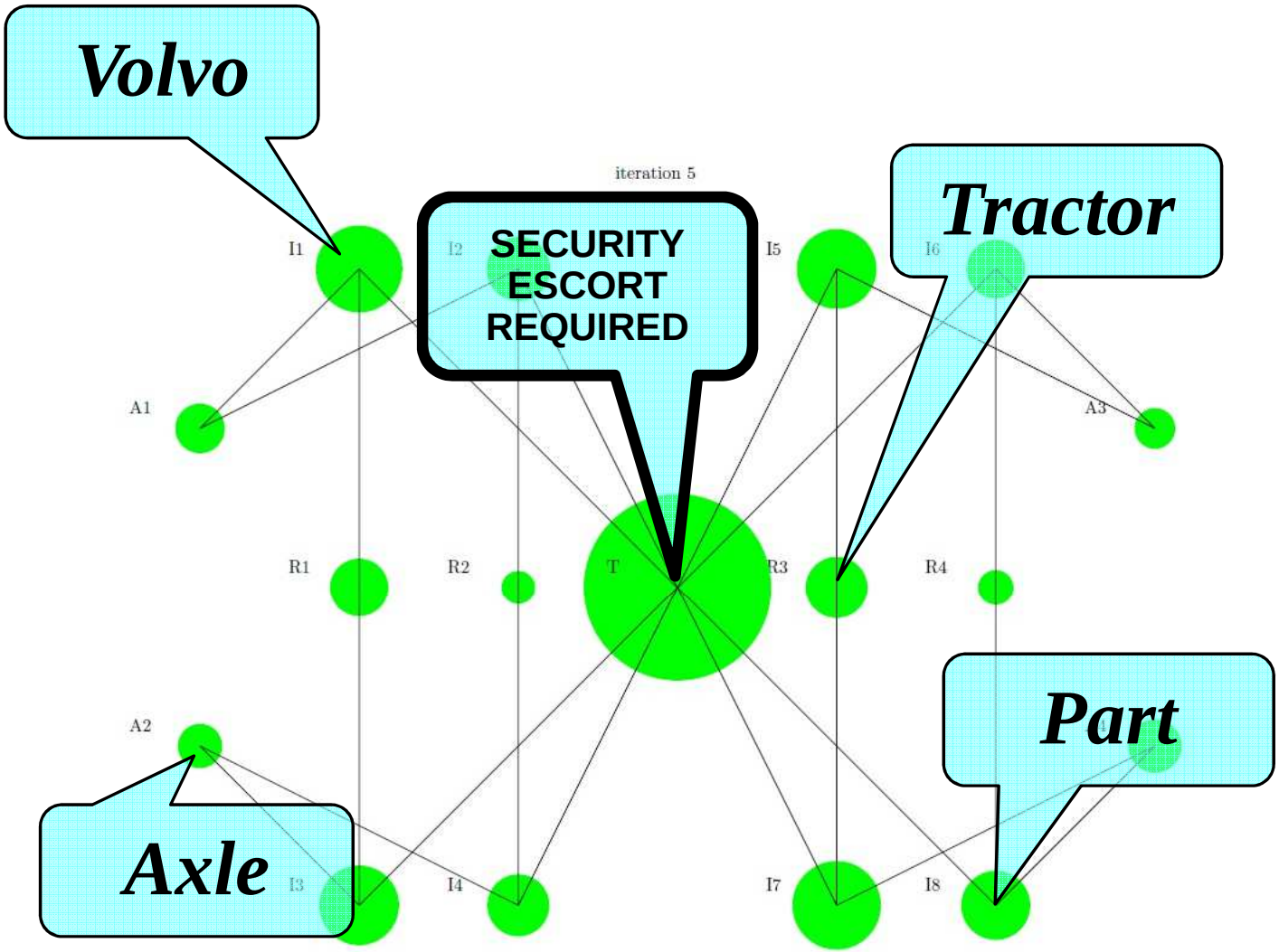

Volvo
iteration 0
SECURITY
ESCORT
REQUIRED
Tractor
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
Part
Axle
I3
I4
I7
I8

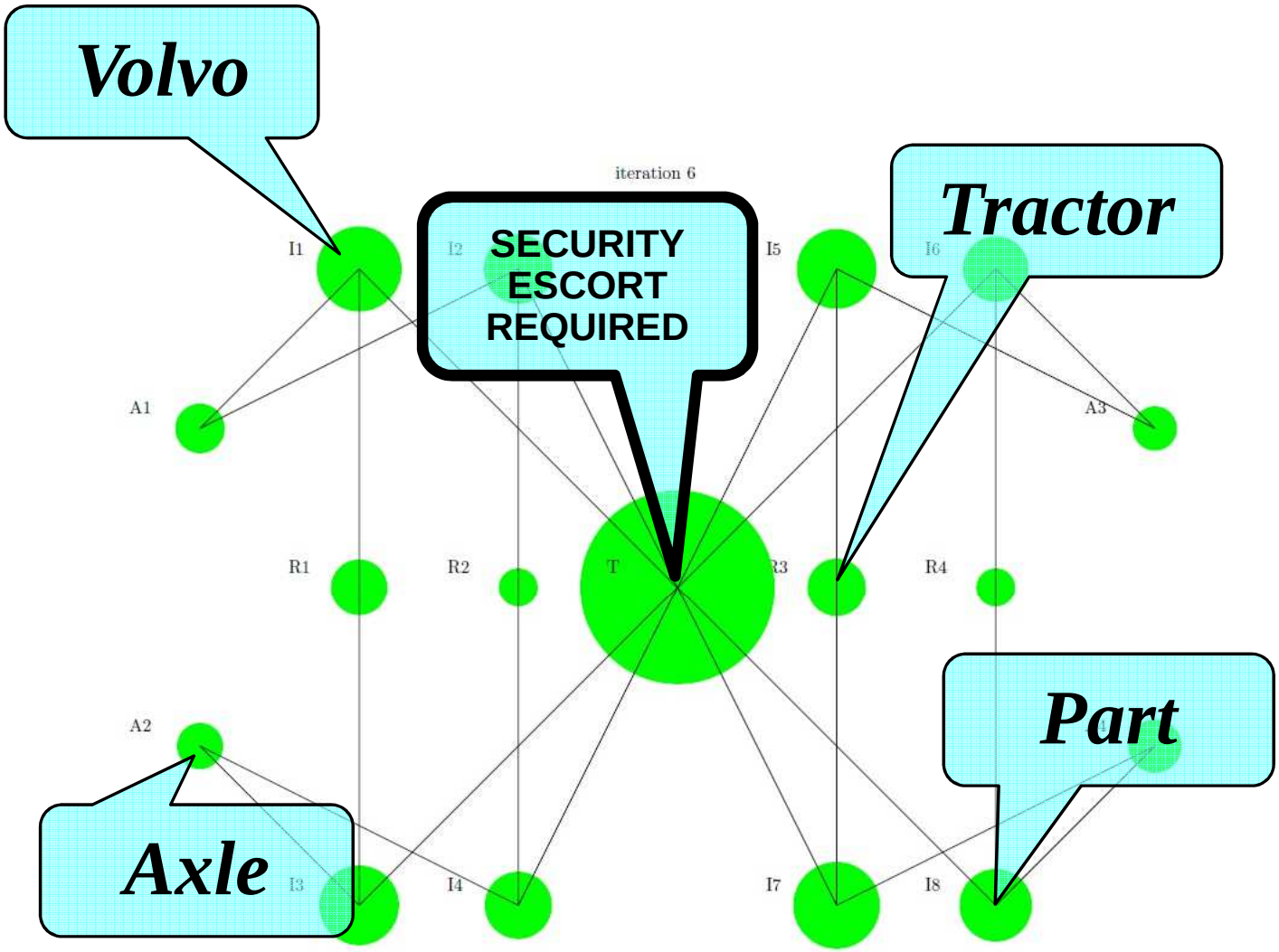

Volvo
iteration 1
SECURITY
ESCORT
REQUIRED
Tractor
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
Part
Axle
I3
I4
I7
I8


Volvo
iteration 2
SECURITY
ESCORT
REQUIRED
Tractor
Part
Axle
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
I3
I4
I7
I8


Volvo
iteration 3
SECURITY
ESCORT
REQUIRED
Tractor
Part
Axle
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
I3
I4
I7
I8


Volvo
iteration 4
SECURITY
ESCORT
REQUIRED
Tractor
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
Part
Axle
I3
I4
I7
I8


Volvo
iteration 5
SECURITY
ESCORT
REQUIRED
Tractor
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
Part
Axle
I3
I4
I7
I8


Volvo
iteration 6
SECURITY
ESCORT
REQUIRED
Tractor
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
Part
Axle
I3
I4
I7
I8

ANOTHER USE CASE OF OUR GRAPH MINING ALGORITHM
Multidimensional multiresolute clustering of massive socio-semantic networks

Multidimensional multiresolute clustering

- The algorithm has been validated on the task of clustering 46 million users of the largest European online social networking service VK, originally VKontakte

- Crawling
  - Crawling has been done through the official API of the VK social network. The limitations and low performance of this API has been largely overcome by tools, which are outside of the scope of this paper. The collected data include:
    - Profiles,
    - Lists of friends,
    - Text posts,
    - Social links between users.

Multidimensional multiresolute clustering

- Crawling has been performed in the period 1st of the August 2016 till 2nd of October 2016 by 25 high performance virtual servers. Technical tools used in multilink flexible structure Web crawler, data collection and analytic module include GreenPlum Database (GPDB) from Pivotal, JSONB, HTTP REST/JSON web-service, Nginx, Python3, DigitalOcean, Python scripts and libraries numpy, scipy, graph-tool, sklearn, xgboost.

- Data curation
  - On this stage we selected out of 320 millions of profiles only those, which look like profiles of real people from Russian Federation who are active on VK, and whose interests could be detected based on their posts, not on what they claim as their interests, since most of the announced interests are not reliable or could not be interpreted. Examples of such claimed interests include profile of a young lady “My interests are friends, brother and HIM!”, and the interest “It is good that I moved to St.Petersburg”. We also discarded profiles where names and family names could not belong and could not be transformed to names and surnames or real people.

Multidimensional multiresolute clustering

- Since many users registered using non-canonical or deliberately alternated forms of their name, we applied a recurrent neural network to check if names and surnames could be normalized to the form, which could belong to real people. This task has been performed based on the linguistic dictionary, the data base of names and surnames of 4 millions real people, the data base of one thousand pairs of transformation: string to real name.

- The performance of the clustering algorithm introduced in this paper is low degree polynomial; algorithm computed interpretable overlapping clustering in 4 days on two servers. The results are presented in several ways, including list of interests sorted by the number of users having this interests, the graph of connected interests, the overlapping clustering of users. All the operation used are “knowledge free”, parameters of modelling and algorithm include selection of several thresholds, most notably, thresholds for term frequencies.

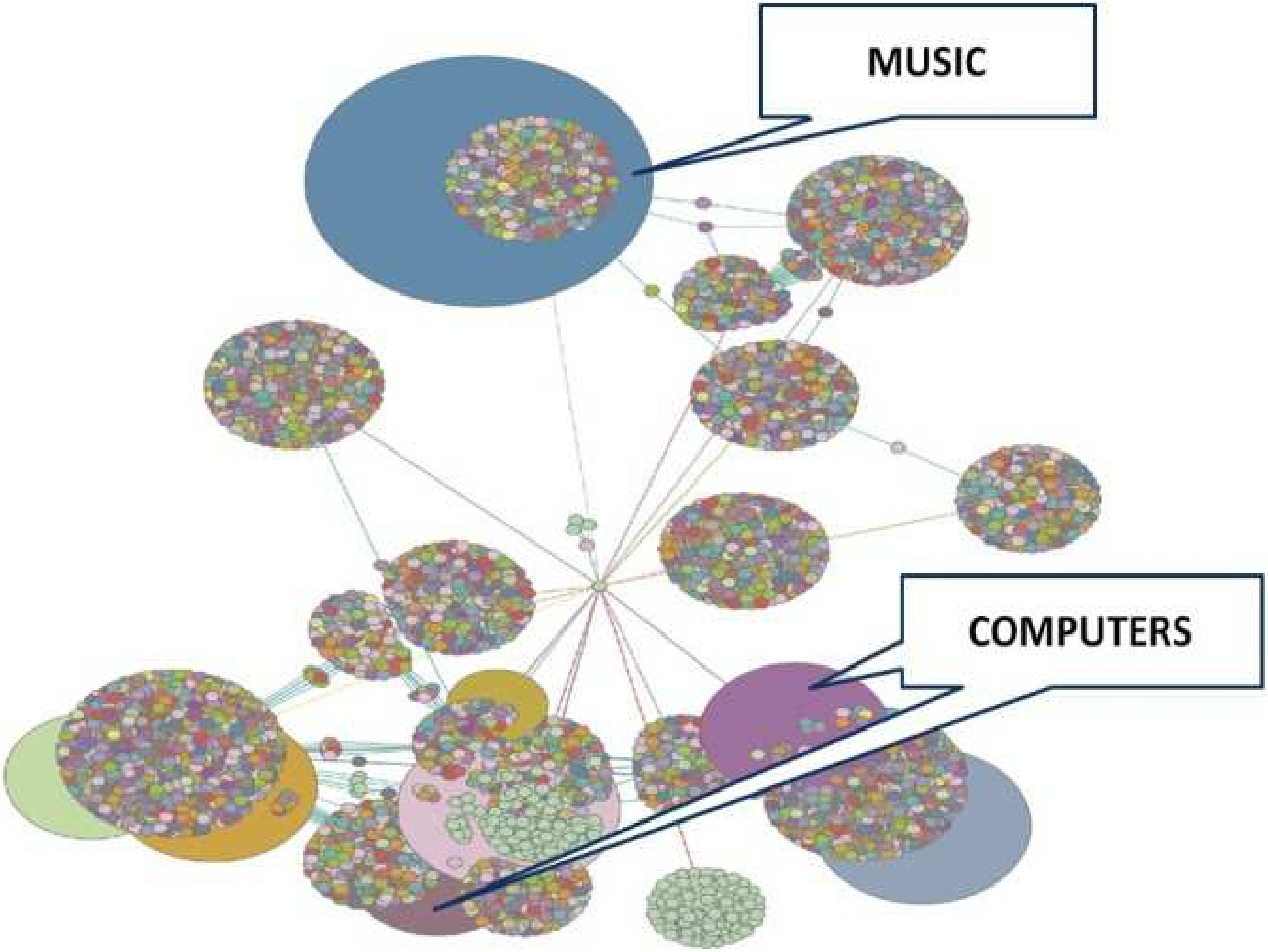
Multidimensional multiresolute clustering

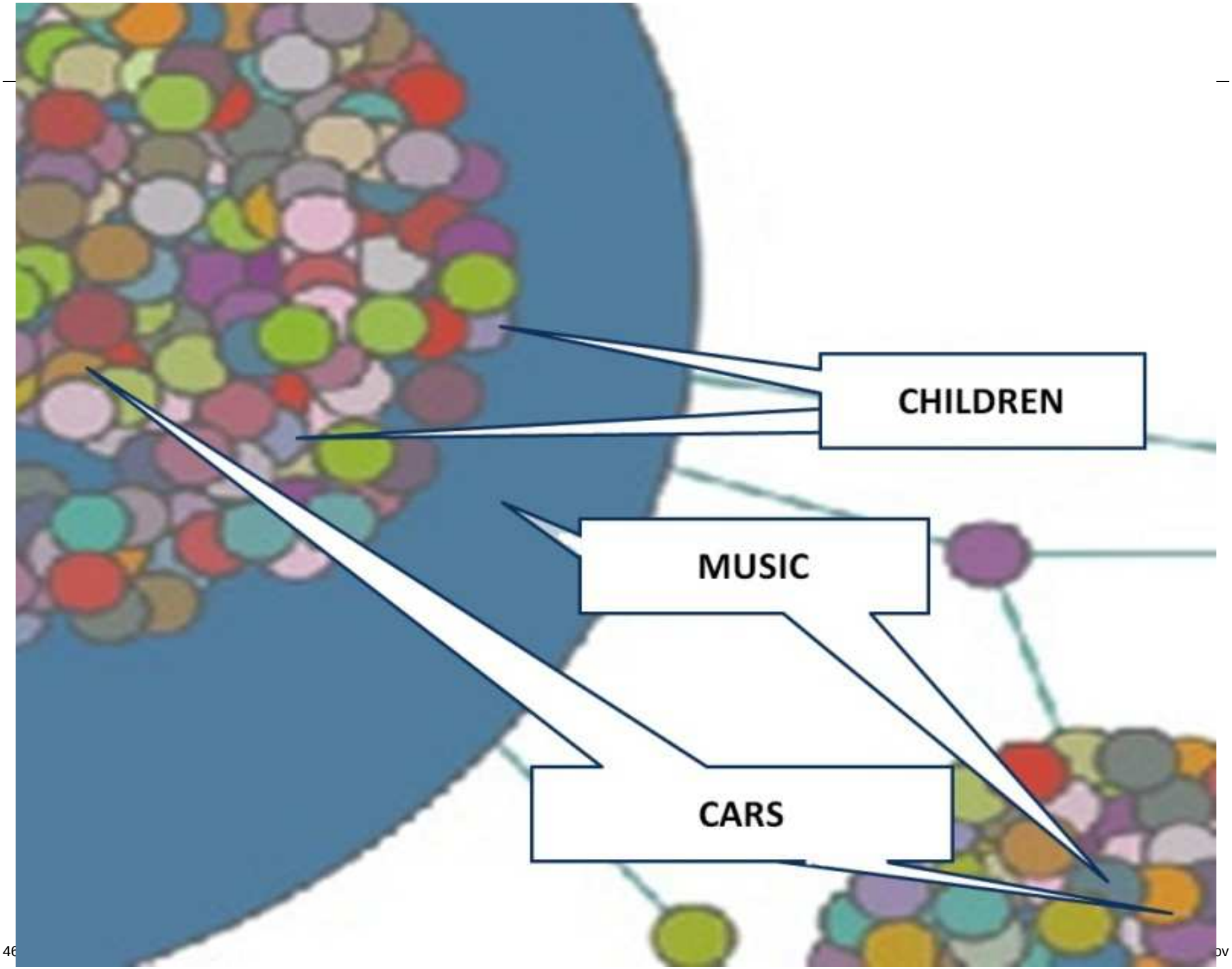
- Multidimensional –
  - we automatically detect the interests of users based on the results of textual analysis of their posts.
  - We automatically label each interest by a set of automatically detected keywords
  - This allows us
    - To give various dimensions, to contextualize social links. To overcome limitations of Friends of Friends notion

      Friend of a friend (FOAF) is a phrase used to refer to someone that one does not know well. The rise of social network services has led to increased use of this term. “Six degrees of separation” and the "small world" phenomenon are related terms.
    - *You are my soccer pal, but your mathematical friends are all boring*
- Cluster – a group of users
  - who all share a particular common interest (in addition to other interests)
  - Any two member of the cluster are connected by a chain of social links within the cluster

The most popular interests of Russian internet users

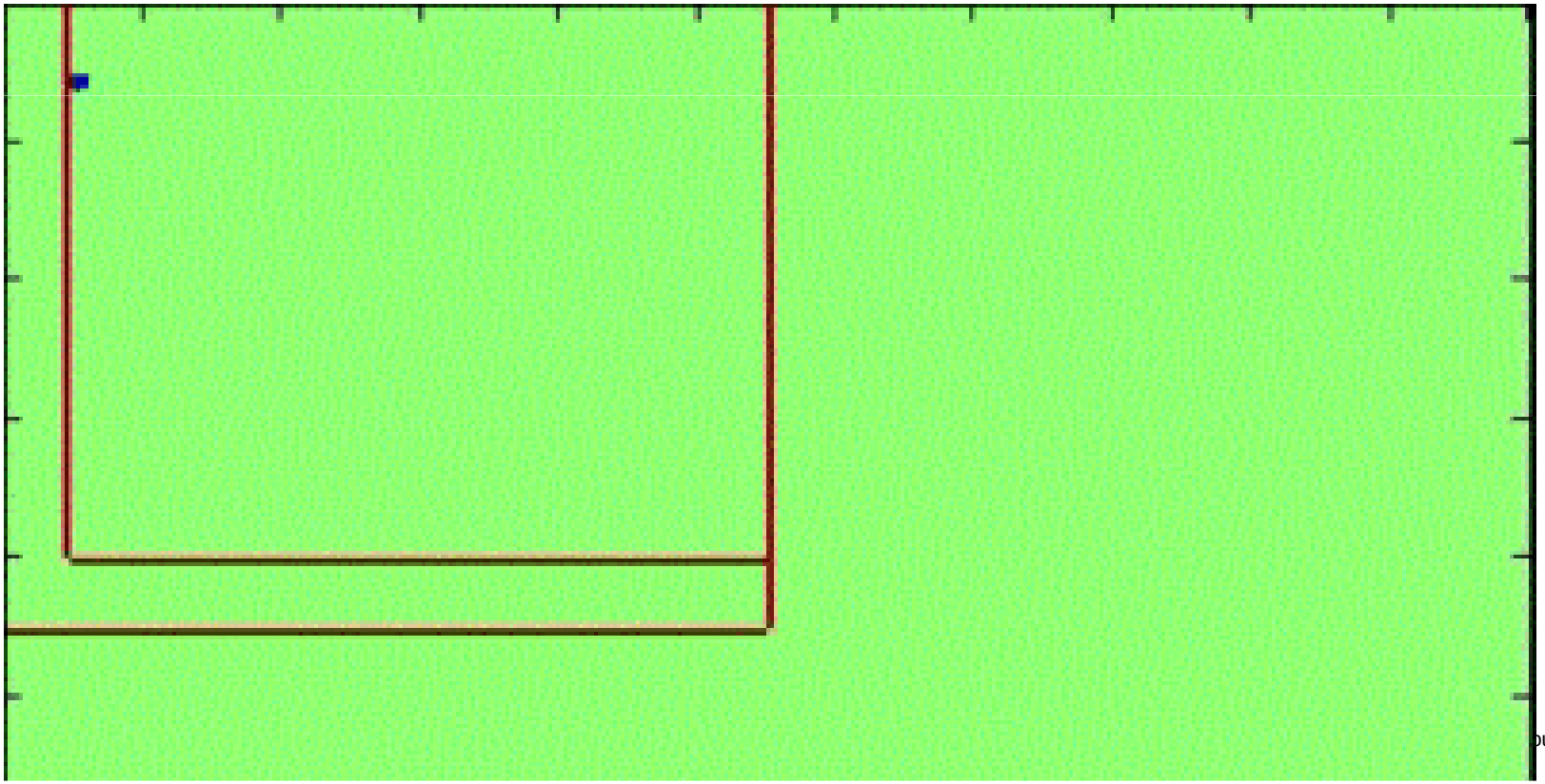
- the most populated groups
  are *music, computers, children*, etc.,
- Most of the user profiles falls simultaneously into several interest categories.
  - The top ten most populated intersections of interests in decreasing of popularity order are
    {recreation, outdoor activity, nature} {travel, children, internet}, {skiing, music, mountain skiing, foreign languages}, {vacations, people, outdoor activities}, {cars, snowboarding, business}, {computers, cars}, {soccer, vacations, cars}, {basketball, cars, business}, {cars, sport}, {psychology, arts}.
    In top 50 one can find the following combinations in decreasing order:
    {fishing, sport, music},{fishing, skiing, nature},{sex, internet, sport, business},{cars, girls},{dancing, theater, fitness, sport, work},{computers, games, computer games},{billiards, cats}, {photography, travels, summer, life, music, family},{photography, cinema, architecture, photo, music, design}
- Multiresolute – the algorithms produces in one go clusters on all levels (micro-, mezzo, macro-level), as well as relations between clusters.
  Topology of clusters is not dendroidal, but dendroidal-like visualisation is useful.


MUSIC
COMPUTERS

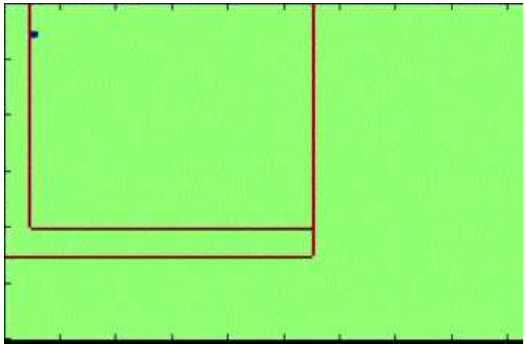

CHILDREN
MUSIC
CARS

Multiscale Structure of the Media (continuous media) is frequently discovered by *Processes*

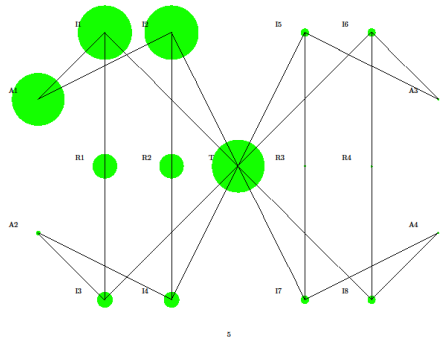
Processes on Networks:
How we study the Earth?
It is good to have master equation (?)
The story of Catenary



Processes in Networks



- How we study the structure of the Earth?
  - RADARS (on aircrafts, ships…) induces processes, The smaller the wave length – the better the resolution
  - THE INTERIOUR OF THE EARTH:
    Using opportunities (Earthquakes, Nuclear tests, …) or creating generating signals and measuring the input signal in one or more points.
- NETWORKS Similarly, the networks are frequently studied by network flow methods
  - introducing the processes where something is flowing from node to node across the edges
  - Why FLOWING? Probably, the interaction could be much more complex



Processes in Networks

- NETWORKS – many massive real life networks, network models of techno-social systems
  SHOULD BE CONSIDERED
  - As massive random networks
  - As discrete (static) models of continuous media
- The structure should be discovered by processes evolving in (discrete) time based on local interactions in the absence of long range forces
- There are problems with this approach
  - Master equation?? – no chance to pick (a good??) the right equation when dealing with a nuanced empirical environment, with contextualized semantics of natural languages and the actions of agents with the free will (next time in exactly the same situation they might behave differently)
  - Work with stencils !? – we don't know micro-forces, don't know the low of interactions, WE DON'T KNOW GEOMETRY AND MECHANICS

Master equation

- Neal Koblitz: "Mathematics as Propaganda“
  Who can argue with an equation? An equation is always exact, indisputable. Challenging someone who can support his claims with an equation is as pointless as arguing with your high school math teacher.
- Johnny Carson was in top form, but the show could have bogged down if his guest had delved into subtleties or overly serious discussion. However, Ehrlich had the perfect solution. He took a piece of posterboard and wrote in large letters for the TV audience:
  D = N * I
  "In this equation," he explained, "D stands for damage to the environment, N stands for the number of people, and I stands for the impact of each person on the environment. This equation shows that the more people, the more pollution. We cannot control pollution without controlling the number of people.“

  Johnny Carson looked at the equation, scratched his head, made a remark about never having been good at math, and commented that it all looked quite impressive.
  ... But what if the viewer is too intimidated by a mathematical equation to apply some common sense?

Processes in Networks

- APPROACH I USED IN RECENT PROJECTS:
  - Work with stencils
  - Synthesize new processes
  - Test solutions, tune parameters
    - Like machine learning
    - Like neural networks
    - Though my approach is actually knowledge based approach which simplify interpretation

Master Equation → Numerical Solution

- Bonacich Power Centrality, Eigenvector Centrality, Google's PageRank

  – "Google's workhorse search engine ranking algorithm, PageRank, is actually a variant on an SNA concept - Bonacich Power Centrality. Bonacich (1987) hypothesized that someone's power in society depends on the power of his or her social contacts. Bonacich formalized this mathematically:

  $$c_i = B(c_1R_{i1} + c_2R_{i2} + \ ... \ + c_nR_{in}) ,$$

  where c_i is the person in question, B is the magnitude of the effect, and R_{ij} is the strength of the relationship between the person in question, i, and each of the other people, j, under consideration.
  If B=1 , the formula becomes eigenvector centrality, of which PageRank is a variant.
  Now, Page, et al. (1998) do not cite Bonacich, I am not claiming that they stole the idea - I am merely stating that a social network analyst appears to me to have been the first to think up the concept".

  Solomon Messing http://www.stanford.edu/~messing/RforSNA.html

Master Equation → Numerical Solution

the eigenvector equation $\mathbf{Ax} = \lambda\mathbf{x}$ → Computation

Master equation easily leads us to a numerical solution

It is great to have “the right master equation”!
What is the shape of a hanging chain?

What is the shape of a hanging chain when supported at its ends and acted on only by its own weight?

- Galileo: “This chain will assume the form of a parabola”

  $y = x^2$

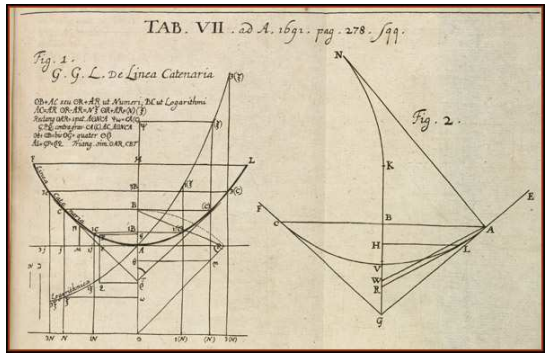
- But the shape is different:

  $y = (a / 2) (e^{x/a} + e^{-x/a})$

  which was established later by applying calculus

  Plotting geometric arrangements and forces acting on small segments of the chain

  Integrating the results

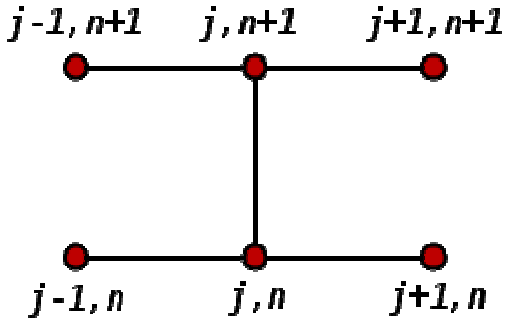


Leibniz's solution is on the left.
Huygen's illustation is on the right.

." In 1669, Jungius disproved Galileo's claim that the curve of a chain hanging under gravity would be a parabola (MacTutor Archive). The curve is also called the alysoid and chainette. The equation was obtained by Leibniz, Huygens, and Johann Bernoulli in 1691 in response to a challenge by Jakob Bernoulli”.

http://mathworld.wolfram.com/Catenary.html

- "Plotting geometric arrangements and forces acting on small segments" evolved into
  - Finite difference method
    - In mathematics, finite-difference methods are numerical methods for approximating the solutions to differential equations using finite difference equations to approximate derivatives.
  - Stencil
    - In mathematics, especially the areas of numerical analysis concentrating on the numerical solution of partial differential equations, a stencil is a geometric arrangement of a nodal group that relate to the point of interest by using a numerical approximation routine. Stencils are the basis for many algorithms to numerically solve partial differential equations.



Numerical Solution → NO Master Equation

- “Integrating” evolved into …
  - Well, in financial mathematics (fast trading) solutions are tuned on “stencils”. Numerical solutions are known. Master equation is not known, and is not interesting to know.
- “Master equation is not known” – this is ok.
  - But we need to be aware about emergency effects in complex systems: learning how to do something right in a small scale, doesn’t necessarily imply that we’ll do right things in a bigger scale

We don't know "geometry" and "mechanics" of the new brave world of big data
and we can't easily guess master equations

- Leibniz, Huygens, and Johann Bernoulli knew geometry and mechanics. We don't know "geometry" and "mechanics" of techno-social systems (and we don't even know "geometry" and "mechanics" of semantic network, social networks, …)
- but we can create small "nodal arrangements" modeling multidimensional networks (for instance, folksonomies)
- Apply known and novel numerical algorithms and utilize state of the art knowledge to decide which algorithms provides better results.
- The next step - to check if good properties of the numerical solutions on the micro-level hold true on the mezzo-level

Mining Customs Declarations for commercial goods

Знание, его представление и моделирование сетями

Способы представления знания

- Есть разные способы: XML, сети, …
- Сети обладают преимуществами в удобстве агрегации разнородной информации
- Сети являются традиционным способом представления знаний, таких как онтологии, отражающие семантические отношения

  Колесо – часть машины. Курица – птица
- Примеры семантических отношений

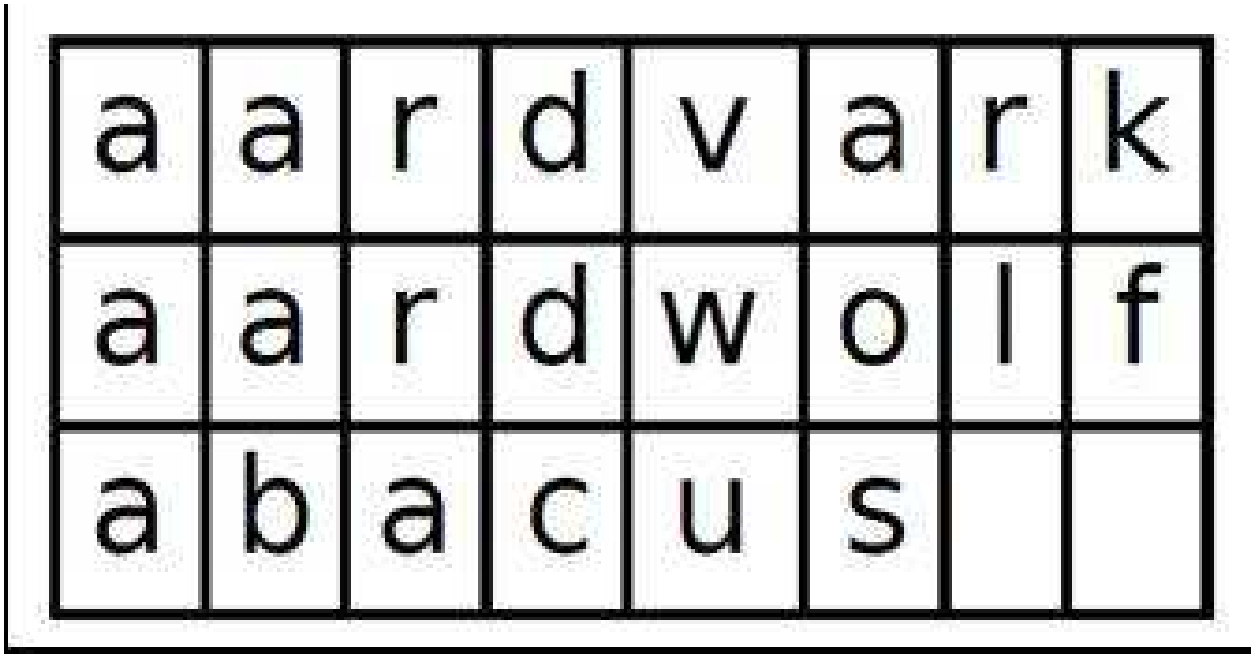
  Меронимия и холонимия, как семантические отношения являются взаимно обратными друг другу. Например, термины *двигатель*, *колесо* и *капот* являются меронимами по отношению к термину *автомобиль*. В свою очередь, термин *автомобиль* является холонимом по отношению к терминам *двигатель*, *колесо* и *капот*.

Способы представления знания (Knowledge Representation)

- Способ моделирования данных должен соответствовать задачам исследования.
- Например, многие данные приходят к нам в табулярной форме, и эта форма может соответствовать целям исследования.
- Но табулярные данные полезно моделировать сетями, как мы проиллюстрируем на следующих слайдах составленных по главе Semantics of Techno-Social Spaces в книге "Modern Computational Models of Semantic Discovery in Natural Language" IGI Global, 2015,
в которой мы подытожили опыт нашей лаборатории по обработке больших данных, содержащих различные специальные коды, сопровождаемые текстовыми описаниями товаров

Network modelling for tabular data

- The rationale of the network modelling for tabular data has been explained in the paper (Maruev et al., 2014b) by analogy with the use of finite state representation in computer processing of morphology in terms of the approach to the modelling and to the veracity-for-purpose, where the purpose is the discovery of hidden patterns in data.
- Many lists of common English words start from the words aardvark, aardwolf, and abacus. For solving crosswords, in might be suitable to model the list in the tabular form like this:



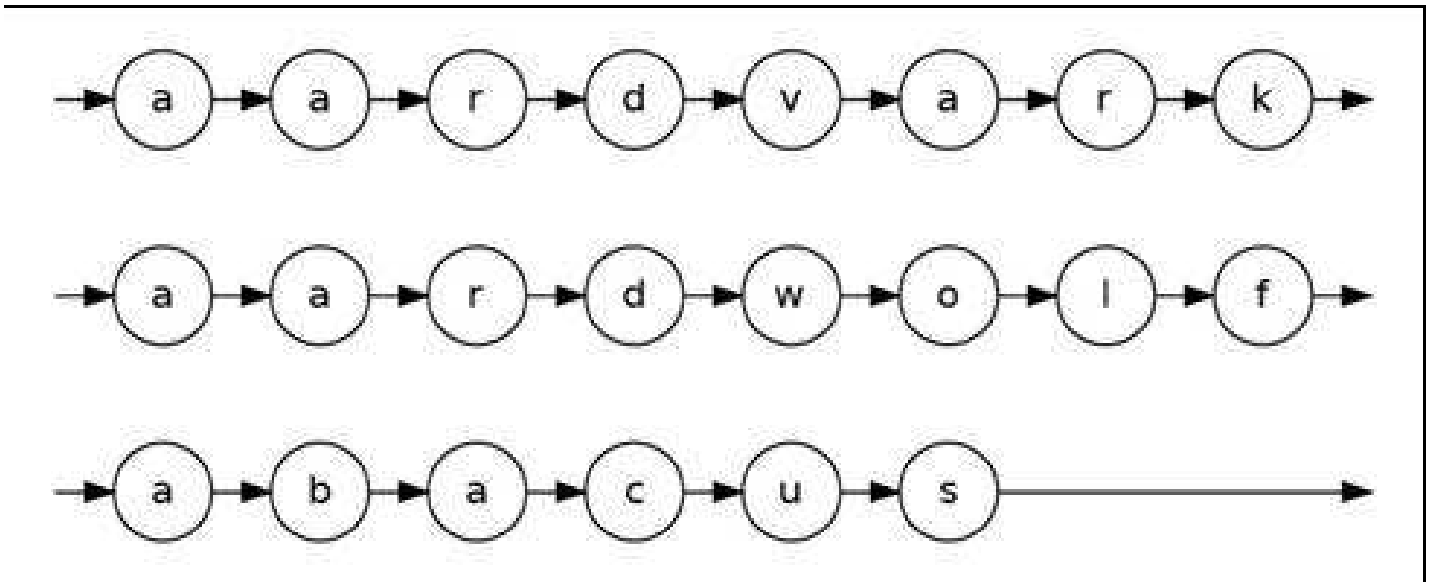
a	a	r	d	v	a	r	k
a	a	r	d	w	o	l	f
a	b	a	c	u	s		

Network modelling for tabular data

- …

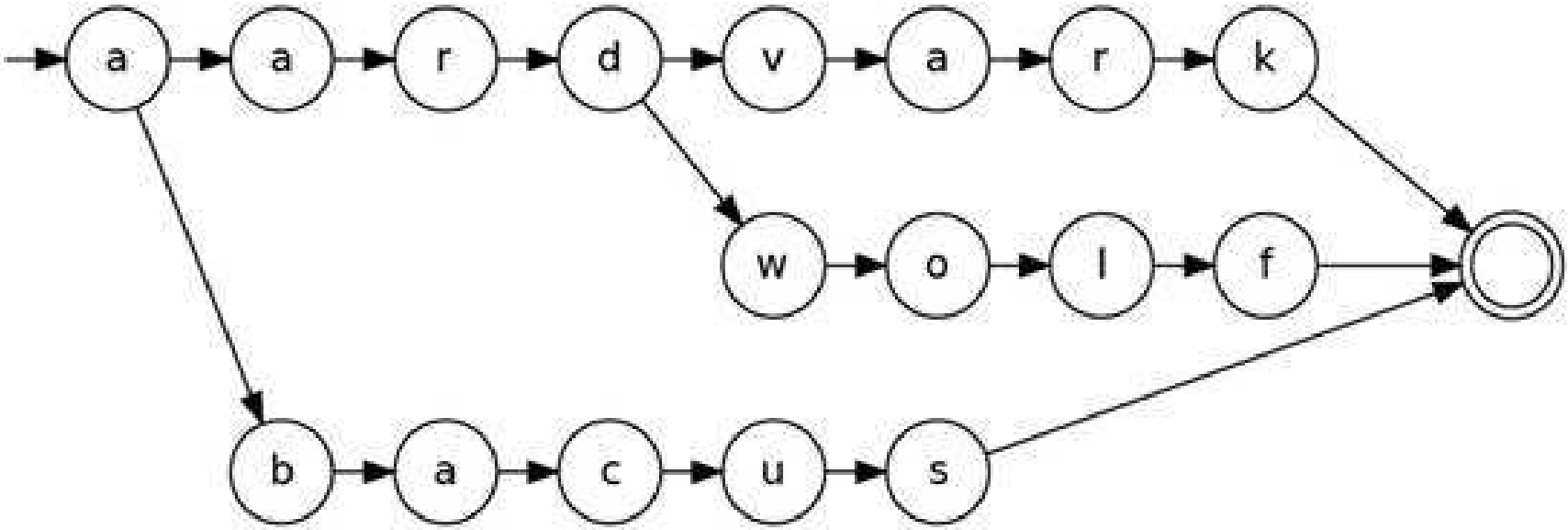
a	a	r	d	v	a	r	k
a	a	r	d	w	o	l	f
a	b	a	c	u	s		

- The same table could be redrawn as a graph where nodes represent letters.



Network modelling for tabular data

- If this graph is used to construct a computer dictionary, it can be compactified to the following form:



Network modelling for tabular data

- When the strings are words from a natural language, graph-based representation of data has at least one crucial advantage over the tabular representation:
  - Firstly, the graph representation becomes very compact since common affixes of words are conflated, and this leads to non-functional advantages (in memory footprint and processing speed).
  - More importantly, even when such compactification is provided by formal mathematical methods, which are unaware of the morphology, effectively they produce a representation of the list of strings in a graph form which shows patterns of the morphology of the language; therefore it becomes possible to process out-of-vocabulary words, like *trichloroisocyanuric*, and to construct morphological guessers (Jurafsky and Martin, 2009). For instance, a morphological guesser might infer that the word *ontologisation* is a well formed English noun, and to find that it is related to the noun ontology. Moreover, using graph-representation one can infer that the relation between the pair of words *ontology-ontologisation* is the same as the relation between words *industry-industrialization* (see, for instance, Troussov, A., & O'Donovan, B., "Morphosyntactic Annotation and Lemmatization Based on the Finite-State Dictionary of Wordformation Elements“, 2003).

- The procedure of the processing out-of-vocabulary words, like the word *ontologisation*, can be summarized in the following diagram:

Input: a new word: *ontologisation* →

→ Network model of existing words →

→ Patterns which the input follows

Network modelling for tabular data

- According to this diagram, a new word, like *ontologisation*, is mapped into the network representing known words. The exact mapping is not possible for out-of-vocabulary words, but morphological parts of the word, including prefixes, roots, and suffixes) will be mapped to corresponding strings of symbols (like *ontolog-* and *-isation*). Graph-based methods are used to provide such mapping and to "to connect the dots" between parts of the of the word *ontologisation*.
- The same diagram has been used for mining of customs declarations. When the network model of customs declaration has been already constructed, the scheme of processing new declarations is the same as it is in the above mentioned morphological applications:

Input: a new customs declaration record →

→ Network model of customs declaration data →

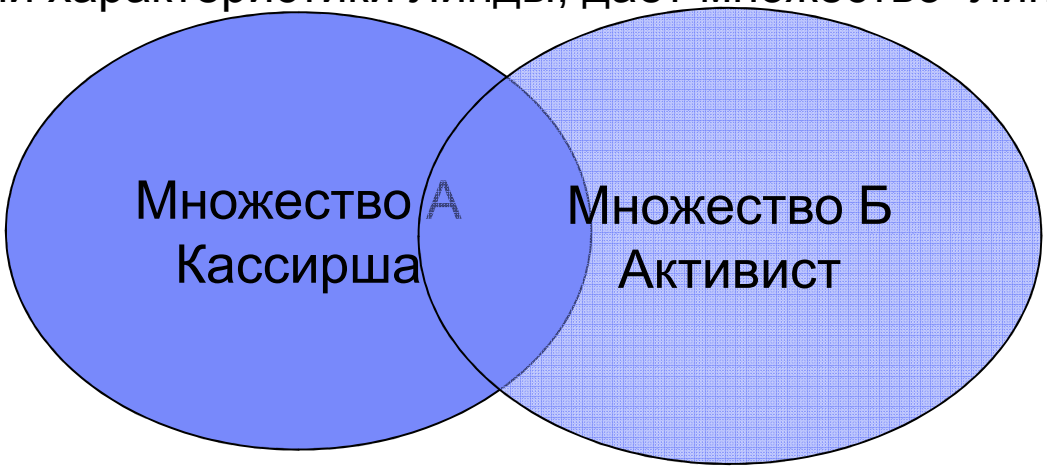
→ Patterns which the input follows

Нечеткая логика

основная идея

Почему нужна в рассматриваемой задаче

- При любом подходе к ответу на вопросы о Чехольчик для Айфона нам придется аггрегировать разные факторы при помощи логических операторов И ИЛИ НЕ … Например, пересечение двух множеств, описывающий характеристики Линды, дает множество Линда – Кассирша и Активист



- Реальные ситуации могут описываться противоречивыми признаками
- Обычная логика ломается, можно переходить к нечеткой

Почему нужна в рассматриваемой задаче

- В нашем случае, комплексные логические операции могут не понадобиться,
- Нам надо только уметь объединить факторы про Чехольчики или про Линду, сделать многокритериальный выбор:
  - Про объект (*чехольчик*) вычислены несколько чисел которые говорят в пользу или против того, что *чехольчик* в данном контексте это *электроника*
  - Надо объединить эти числа в одно, по которому можно давать окончательный ответ (для чего надо сбалансировать *пропуск цели* и *ложные тревоги*)
- Одна из важнейших полярностей в нечеткой логики – это возможность параметров компенсировать друг друга
- В задаче про обработку там. Декл, мы учим эту полярность из данных. Для этого мы
  - Для объединения используем L^p
  - Параметр *p* – учим из данных

L^p norm

Given an n-dimensional vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

a general vector norm $|\mathbf{x}|$, sometimes written with a double bar as $\|\mathbf{x}\|$, is a nonnegative norm defined such that

1. $|\mathbf{x}| > 0$ when $\mathbf{x} \neq \mathbf{0}$ and $|\mathbf{x}| = 0$ iff $\mathbf{x} = \mathbf{0}$.

2. $|k\,\mathbf{x}| = |k|\,|\mathbf{x}|$ for any scalar k.

3. $|\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}|$.

In this work, a single bar is used to denote a vector norm, absolute value, or complex modulus, while a double bar is reserved for denoting a matrix norm.

The vector norm $|\mathbf{x}|_p$ for $p = 1, 2, \ldots$ is defined as

$$|\mathbf{x}|_p \equiv \left(\sum_i |x_i|^p \right)^{1/p}.$$

The p-norm of vector v is implemented as `Norm[v, p]`, with the 2-norm being returned by `Norm[v]`.

The special case $|\mathbf{x}|_\infty$ is defined as

$$|\mathbf{x}|_\infty \equiv \max |x_i|.$$

http://mathworld.wolfram.com/VectorNorm.html

L^p norm

The most commonly encountered vector norm (often simply called "the norm" of a vector, or sometimes the magnitude of a vector) is the L2-norm, given by

$$|\mathbf{x}|_2 = |\mathbf{x}| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}\ .$$

http://mathworld.wolfram.com/VectorNorm.html

This and other types of vector norms are summarized in the following table, together with the value of the norm for the example vector $\mathbf{v} = (1, 2, 3)$.

name	symbol	value	approx.
L^1-norm	$\|\mathbf{x}\|_1$	6	6.000
L^2-norm	$\|\mathbf{x}\|_2$	$\sqrt{14}$	3.742
L^3-norm	$\|\mathbf{x}\|_3$	$6^{2/3}$	3.302
L^4-norm	$\|\mathbf{x}\|_4$	$2^{1/4}\sqrt{7}$	3.146
L^∞-norm	$\|\mathbf{x}\|_\infty$	3	3.000

- Вы ухватываете суть вопроса о компенсации параметров в нечеткой логики если сможете ответить на этот вопрос
- Какой параметр *p* правильнее использовать?…
  - x_i – денежные поступления из разных источников, задача – баланс
  - x_i – надежность атомной злектростанции по фактору *I*; чем больше число, тем надежнее станция. Задача – оценка общей надежности станции

MINING TABULAR DATA

It has been a long, rainy day …

- At a checkpoint between Russian Federation and an EU country a Russian customs officer is inspecting a track with commercial goods going to Russia.

It has been a long, rainy day …

- **At a checkpoint between Russian Federation and an EU country a Russian customs officer is inspecting a track with commercial goods going to Russia.**

Customs Declaration – an example

An example of the itemised document to complete custom declaration for goods transported in one vehicle.

№	DI	E	C1	C2	C3	Goods-TNVEDCode	GoodsDescription	GW	InvoicedCost	CC	CR
7	11	0	967	218	204	9503003500	ИГРУШКИ	159.700	3985.64	EUR	44.0129
8	11	0	967	218	204	2208308200	ВИСКИ	295.950	943.20	EUR	44.0129
9	11	0	967	218	204	4820103000	БЛОКНОТЫ ДЛЯ ЗАПИСЕЙ	15.140	128.64	EUR	44.0129

It has been a long, rainy day …

- Gr

912	30	8486209009	СИСТЕМА КЛАСТЕР ПЛАЗМА ЛАБ 100 С КАМЕР	8326.000	4014349.56	USD
912	30	8465930000	ТОЧНОЕ ПОЛИРОВОЧНОЕ ОБОРУДОВАНИЕ	156.300	61285.00	USD
31	465	8708809909	СТУПИЦА ПЕРЕДНЕГО МОСТА	160.000	80.00	USD
31	465	870899	ПЕРЕДНЯЯ ЧАСТЬ ТЯГАЧА VOLVO FH 12	4400.000	1600.00	USD
31	465	870899	ПЕРЕДНЯЯ ЧАСТЬ ТЯГАЧА MERSEDES BENZ AC	3240.000	1200.00	USD
31	465	870899	ПЕРЕДНЯЯ ЧАСТЬ ТЯГАЧА IVECO STRALIS 5/	4540.000	1640.00	USD
218	204	9503003500	ИГРУШКИ	159.700	3985.64	EUR
218	204	2208308200	ВИСКИ	295.950	943.20	EUR
218	204	9503004100	ИГРУШКИ НАБИВНЫЕ	7.310	207.00	EUR
218	204	4820103000	БЛОКНОТЫ ДЛЯ ЗАПИСЕЙ	15.140	128.64	EUR
218	204	2402209000	СИГАРЕТЫ	2399.570	83950.80	EUR
218	204	2208701000	ЛИКЕР	2791.390	10220.34	EUR
218	204	3923210000	ПАКЕТЫ ПОЛИЭТИЛЕННОВЫЕ	877.500	5100.00	EUR
218	204	9608200000	РУЧКИ	1.860	74.16	EUR
218	204	2205101000	ВЕРМУТ	3663.860	9185.52	EUR
218	204	2208403100	РОМ	7004.870	25627.20	EUR
218	204	9503002100	КУКЛЫ	46.670	1737.95	EUR

People make errors, some people commit fraud

- There are many “hard rules” in customs regulation, but many important decision are left at the discretion of the customs officer.
- AND
  - –People make errors,
  - –Some people commit fraud

- a computer system is able to help

- To a Russian customs officer
  - By raising red flags about potential discrepancies in declarations
  - By recommending actions (like assigning security convoy vehicle for a truck transporting goods to the Russian Federation)
- To a Dutch trader, Polish carrier, Swiss insurer by indicating potential problems with the customs declaration for goods to Russian Federation

MINING TABULAR DATA

- **Pilot project run in collaboration with the Federal Customs Service of Russia and Irish Office of the Revenue Commissioners to create proof-of-the-concept scalable technology platform for mining tabular data**
- **Particularly to create proof-of-the-concept prototype for mining custom declarations for cases not covered by hard rules of custom regulation**
  - Cold start: capable of finding patterns even in small amount of row data, test new records if they follow certain patterns
  - Ease of injection of external knowledge (for instance, relations between customs codes, models of misspellings)
  - Provide recommendation based on found patterns
  - Provide explanations of the rationale behind such recommendations

Tabular representation of customs declarations

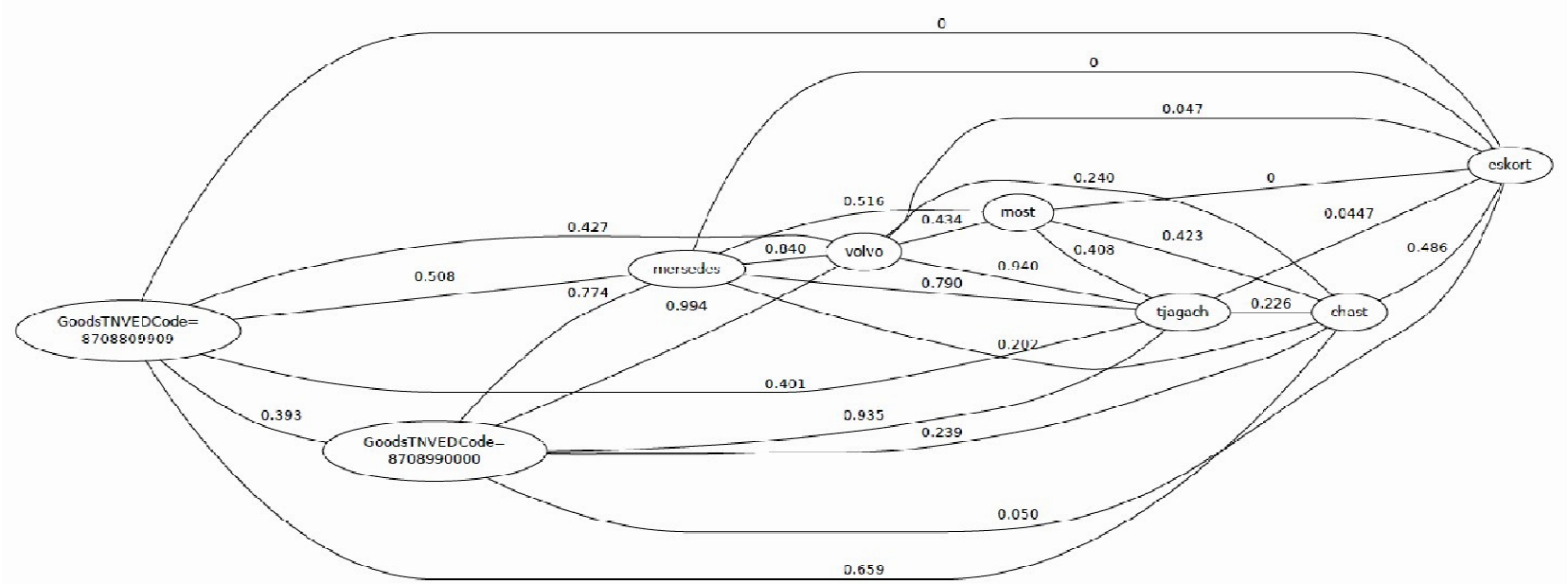
912	30	8486209009	СИСТЕМА КЛАСТЕР ПЛАЗМА ЛАБ 100 С КАМЕР	8326.000	4014349.56	USD
912	30	8465930000	ТОЧНОЕ ПОЛИРОВОЧНОЕ ОБОРУДОВАНИЕ	156.300	61285.00	USD
31	465	8708809909	СТУПИЦА ПЕРЕДНЕГО МОСТА	160.000	80.00	USD
31	465	870899	ПЕРЕДНЯЯ ЧАСТЬ ТЯГАЧА VOLVO FH 12	4400.000	1600.00	USD
31	465	870899	ПЕРЕДНЯЯ ЧАСТЬ ТЯГАЧА MERSEDES BENZ AC	3240.000	1200.00	USD
31	465	870899	ПЕРЕДНЯЯ ЧАСТЬ ТЯГАЧА IVECO STRALIS 5	4540.000	1640.00	USD
218	204	9503003500	ИГРУШКИ	159.700	3985.64	EUR
218	204	2208308200	ВИСКИ	295.950	943.20	EUR
218	204	9503004100	ИГРУШКИ НАБИВНЫЕ	7.310	207.00	EUR
218	204	4820103000	БЛОКНОТЫ ДЛЯ ЗАПИСЕЙ	15.140	128.64	EUR
218	204	2402209000	СИГАРЕТЫ	2399.570	83950.80	EUR
218	204	2208701000	ЛИКЕР	2791.390	10220.34	EUR
218	204	3923210000	ПАКЕТЫ ПОЛИЭТИЛЕННОВЫЕ	877.500	5100.00	EUR
218	204	9608200000	РУЧКИ	1.860	74.16	EUR
218	204	2205101000	ВЕРМУТ	3663.860	9185.52	EUR
218	204	2208403100	РОМ	7004.870	25627.20	EUR
218	204	9503002100	КУКЛЫ	46.670	1737.95	EUR

DATE MODELLING

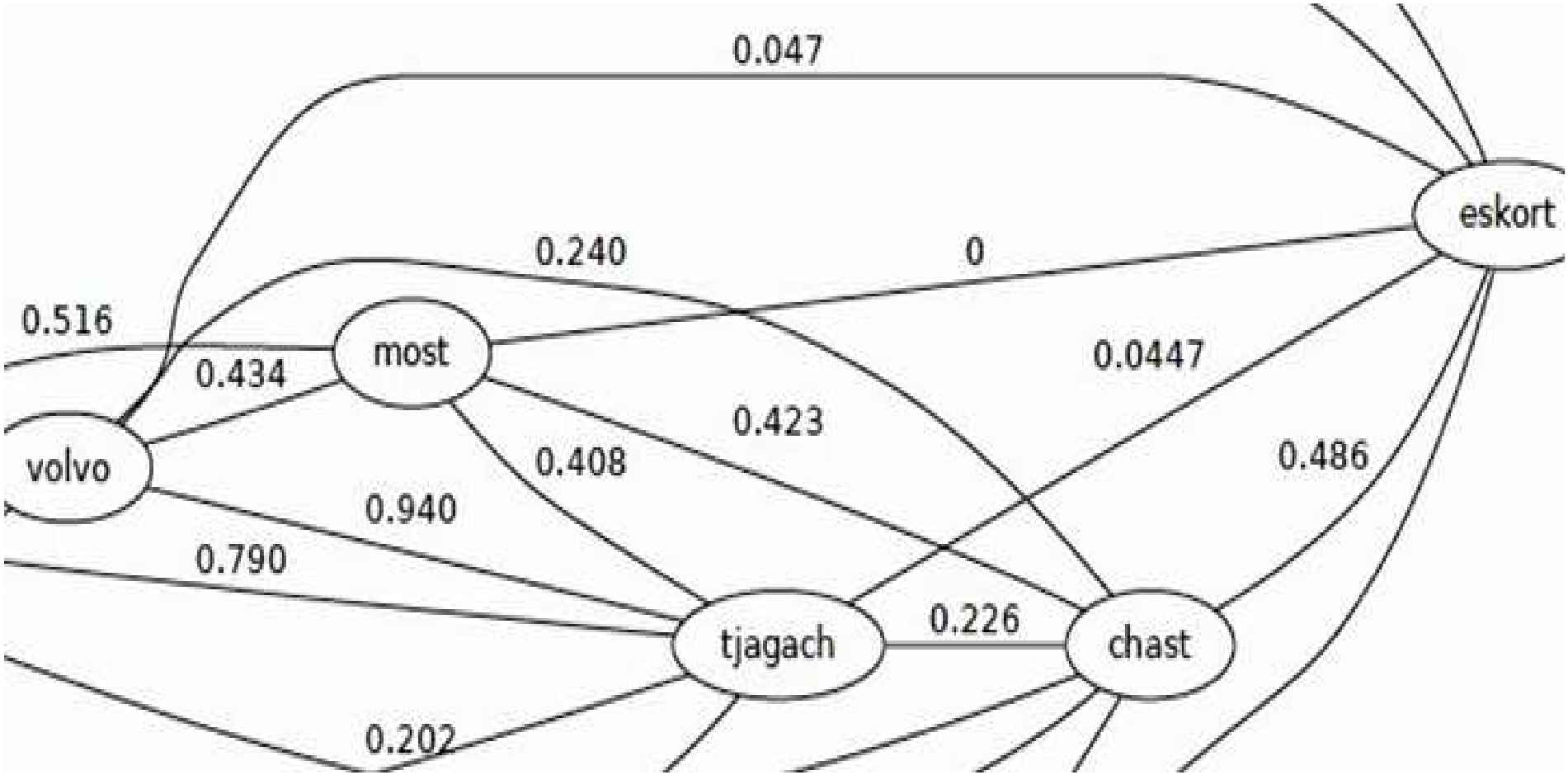
Modelling by multidimensional network

- *Multidimensional …*
  - *– having or involving or marked by several dimensions or aspects*
  - *–… A multi-dimensional database is structured by a combination of data from various sources*
- Multidimensional Network - nodes represent
  - –various abstract codes, for instance, assignment of the security escort, Customs Codes (Commodity codes, GoodsNTVEDCodes)
  - –numerical values measured in kg, seconds, USD etc,
  - –Words, natural language descriptions
  - –Actors (consignee, consignors, carriers)

Modelling by multidimensional network



- The fragment of the network which represents the data from custom declarations.
- Entities (Cells of the table shown before), after preprocessing are merged into one network, where nodes represent words, abstract codes (like assignment of the security escort)
- If two entities are met at least once in the same shipment document, the corresponding pair of nodes is connected by an arc. The weight of that arc represents how frequently the two entities are met in shipment documents (i.e the number of co-occurrences divided by the number of items)



- The fragment of the network which represents the data from custom declarations. The are two types of nodes on this figure: words and security escort tag.
- This network shows, for example, that the word “tjagach” (“тягач” in Russian; rooter, tractor in English) was met in good descriptions which require armed escort in 0.0447% cases. The shipments which has in the description the word “mersedes” (“мерседес” in Russian; Mercedes-Benz international) never required armed escort, while “volvo” required escort in 0,047% of cases. Shipments with parts, details of something (“chast”, “часть” in Russian), frequently were escorted.

MINING

Scheme of applications

- Network – is a representaion of the knowledge about the past
- Knowledge allows us make “inferencing” for new data:
  - –We project new data onto the network
    i.e. onto the knowledge about the past
  - –and investigate consequences
- We can explain this mining in mathematical terms as the transformation of functions on the nodes of the network:

Novel graph mining method

- We show that a wide range of graph-based algorithms popular in various application domains use the idea of propagation between nodes
- We discuss drawbacks and limitations of these algorithms belonging to the class of network flow algorithms (Borgatti 2005)
- To overcome these limitations, we are developing a much broader class of “physics-inspired” algorithms where the interaction between neighbor nodes could not always be interpreted as a flow producing a diffusion like process.
  - we show that iterative computational schemes similar to those used in network flow algorithms have been used for a long time in finite element analysis to solve physical problems and discuss a class of “physics-inspired” algorithms
- And show the road map to combine physics-inspired algorithms with “logic-inspired” algorithms on graph based on the extension of cellular automata procedure that determines the new state of a cell for the next generation
- We show the applications of these novel class of algorithms to core tasks of network mining – centrality measurements and clustering

PROPAGATION ALGORITHMS AND THEIR USE

- Formally, solution of many network data mining tasks boils down to the following problem: Given an initial function $F_0(v)$ on the network nodes, construct the function $F_{lim}(v)$ which provides the answer.
  - In different domains the function F_0 could be referred to as the initial conditions, the initial activation, semantic model of a text, etc.
  - In ontology based text processing, the initial function F_0 is the semantic model of a text w.r.t. to the knowledge: for instance, $F_0(v)=0$ if the concept v is not mentioned in the text, $F_0(v)=n$ if the concept v is mentioned n times.
    The function $F_{lim}(v)$ should show the foci of the text; for instance, $Argmax(F_{lim})$ is the most important focus of the text, while $F_{lim}(Argmax(F_{lim}))$ is the numerical value of the "relevancy"
  - In IR, the link analysis (such as Google's PageRank) ranks web pages based on the global topology of the network by computing $F_{lim}(v)$ using the iterative procedure where the initial condition is that all web pages are equally "important" ($F_0(v)\equiv 1$),

PROPAGATION ALGORITHMS AND THEIR USE (Cont.)

- Computationally efficient and scalable algorithms usually compute the function F_{lim} (which provides the "answer") using iterations: on each iteration the value of $F_{n+1}(v)$ is computed depending on the values of the function F_n on the nodes connected to the node v
  - Very broad range of algorithms including Google's PageRank, spreading activation, computation of eigenvector centrality using the adjacency matrix.
- Most of the mathematical algorithms behind such iterative computations are propagation algorithms (the "network flow" algorithms):
  they are based on the idea that something is flowing between the nodes across the links, and the structural prominence of nodes could be explained and computed in terms of incoming, outcoming and passing through traffic
- Similar iterative computational schemes have been used for long time in finite element analysis to solve physical problems including propagation of heat, of mechanical tensions, oscillations, etc.
  - Although finite element analysis automata usually perform on rectangular (cubic, etc.) grids, the extension to arbitrary networks is feasible.

PROPAGATION ALGORITHMS AND THEIR USE (Cont.)

- However, the interaction between the material points in mechanics could not always be described as a flow, and such interactions could model more complex processes than diffusion
  - For instance, one dimensional heat transfer equations can be numerically simulated on a one-dimensional mesh by iterations. On each iteration recomputation is based on the formula below:

    $$F_{new}(v) = (F(RightNeighbour(v)) + F(LeftNeighbour(v))) / 2$$

    This linear equation confirms the perception of the heat transfer as a flow: on each iteration the heat – the value of the function F – flows from nodes to the neighbour nodes.

    In physics, a conservation law states that the amount of heat in an isolated physical system does not change as the system evolves, so “move mechanism” in heat propagation is a transfer. In network theory applications, network flow could be also done by “copy mechanism”, such as replication in spread of deceases.
  - At the same time, in physics, many processes can not be interpreted as a flow and can not be described by a function of one real variable. For instance, to simulate the behavior of an oscillating string one needs to operate with three values at each node - position, mass and velocity of the material point corresponding to the node. And none of these properties “flow” to the neighbors.

VALIDATION

- TWO USE CASES WHICH WERE IDENTIFIED BY EXPERTS IN THE FIELD AS PRACTICALLY IMPORTANT
  - checking if the textual description of goods is consistent with the other parameters of the declaration,
  - and assigning of a security escort vehicle for a truck transporting goods to the Russian Federation.
- Validation of the approach has been carried out on the real world data that were not used in building the models; in both scenarios the achieved accuracy was 100 percent in most important use cases.
- The lowest reported accuracy was 90.5 percent, but the data used were insufficient for the completion of the task (the data had no time stamps, while customs codes for fruits imported to Russian Federation, such as tomato, frequently depend on the season).
  However, the accuracy measured in top two results, was 100 percent.

Injection of Fuzzy Logic

- Our algorithm exploits generic computational scheme on graphs as it has been described in Troussov et al. Spreading Activation Methods. In Shawkat A., Xiang, Y. (eds). Dynamic and Advanced Data Mining for Progressing Technological Development, IGI Global, USA, 2009
- However, in this paper the fuzzy logic has been injected into this scheme; that is the affect that neighbour nodes produce on a node is considered as a logical operation AND.
  - The logical operation has a parameter which allows to regulate how well parameters of operation can compensate each other. To explain this phrase one can use the direct analogy between fuzzy logic operation AND and the operation used in arithmetic to compute means of several numbers. The case of full compensation corresponds to arithmetic mean, which is if one of the arguments will be increased by a certain value, and the other will be decreased by the same value, the results will be the same. However, in very many practical applications arithmetic mean is not an appropriate method for calculating an average, and other methods, including geometric mean, are used.

Fuzzy Logic: The experimental results strongly confirmed that:

- Graph mining using the generic spreading activation algorithm described in [Troussov et al., 2009] does not provide good results;
- Injection of fuzzy logic makes the spreading activation method, which is a wide class of algorithms, suitable for both cases;
- The two use cases considered in this paper are “the polar” use cases in terms of the essential properties of the algorithms providing accurate prediction.
  - In the task of security escort assigning, most reliable features should dominate the results trumping less reliable predictors; which in terms of fuzzy logic operations could be expressed as follows: the results of the operation AND should be highly skewed towards the value of the bigger operand.
    The logical aggregation skewed towards the value of the bigger operand spectacularly fails for the other use case, which is the prediction the nomenclature code of goods based on the words in the textual description. Authors demonstrated that for this use case, significant number of weak predictors easily overrules the judgment based on strong predictors.
- There is a one parametric family of logical AND operations, which covers both polar use cases discussed above. The value of the parameter which provides the best solution for a particular use case can be automatically learned from the data (and the formal description of the task modelled by subsets of nodes).

Semantics in a nutshell (word senses, ontologies, compositionality)

Semantics in a nutshell (word senses, ontologies, compositionality)

- *"The apple is violet"* might be a wrong assertion, but a correct answer at an optical test.
- Terms like *Smartphones* or *iPhone 5* are related to electronic goods, and when met in customs declarations for commercial goods might entail premium customs duties and the assignment of a security escort for the shipment.
- Textual descriptions of goods in customs declarations, as well as many other texts on the web, are short and are not always grammatically correct; therefore it might be seen as rather unexpected, that a rudimentary (in terms of the use of linguistic tools and linguistic data, such as ontologies) text processing can easily derive from data that
  - terms like *"Cases for iPhones"* have essentially nothing to do with *electronics* when it concerns the functioning of the international trade.
  - While *"Bullets for Guns"* require the same attention as *Guns*
- To stress why this is surprising, we need to point out that the expression *"Cases for iPhones"* is semantically compositional (that is, the meaning of this expression is determined mostly by its constituents), and syntactically stands in the same relation *N Prep N* to *iPhones*, as the expression *"Bullets for Guns"* stands to *Guns*. The reason of that success is that semantics of words or expressions in this paper has been derived from the data, and has been expressed in the form of relations between terms and other artefacts of the system (including customs duties and assignment of a security escort for the shipment).

Semantics of multidimensional spaces

Semantics of multidimensional spaces

- Frege's context principle states that words have meaning only in the context of a sentence.
  - It is widely agreed that one of the most important principles in the philosophy of language is Frege's context principle, which states that words have meaning only in the context of a sentence, that a philosopher should "never ... ask for the meaning of a word in isolation, but only in the context of a proposition" (Frege, 1884). The context principle also figures prominently in the work of Bertrand Russell and Ludwig Wittgenstein
- We put forward the hypothesis that semantics of the textual content in multidimensional data (like tabular data) should be also analysed in the context of the whole system.
- The hypothesis is substantiated by the introduction of a method of formal modelling of tabular data
- and use of graph-based methods for mining such models,
- and demonstrate applicability of our method for traditional tasks of natural language processing (including term disambiguation and finding semantic foci of a document), as well as for applications in various recommender systems based on a hybrid approach when textual analysis is combined with link analysis.

Semantics of multidimensional spaces: CONCLUSION

- *Frege's context principle:*
  "never ... ask for the meaning of a word in isolation, but only in the context of a proposition" (Frege, 1884).

- *We put forward hypotheses:*
  in multidimensional spaces (like the set of customs declarations), never ask for the meaning of a word like "iPhone", but only in the context of all relations of words and other artifacts (including abstract codes and actors) of the multidimensional space
  For instance, "Cases for iPhones" have essentially nothing to do with electronics when it concerns the functioning of the international trade.

Semantics of multidimensional spaces: CONCLUSION

- *But "Bullets for Guns" require the same attention as "Guns"*

Thank you!

Александр Васильевич Трусов

Alexander Troussov
troussov@gmail.com