Context Information and User Profiling

Marek Kumpošt

Laboratory of Security and Applied Cryptography (LaBAK)

Faculty of Informatics
Masaryk University
Brno, Czech Republic

Contents

Content

Introduction

- Customizable services
  - Operate with "user profiles"
  - Reflects user's previous behaviour (based on their context information)
- Context information
  - Descriptive type of information
  - By-product of on-line activity, associated with an individual
  - May reveal some private information
- User behaviour model (context model)
  - User profiling based on previous behaviour (context)
- Representative behavioral patterns
  - Identification of groups with the same behavioral characteristics
  - Try to identify user(s) by using their behavioral patterns only
- Impact on **users' privacy** (ISPs have huge traffic databases available)
- Techniques for finding behavioral characteristics
  - Input data restriction and optimization
  - Processing data (appropriate input information; data mining techniques)
  - Results evaluation → impacts on users' privacy

Introduction

- Customizable services
  - Operate with "user profiles"
  - Reflects user's previous behaviour (based on their context information)
- Context information
  - Descriptive type of information
  - By-product of on-line activity, associated with an individual
  - May reveal some private information
- User behaviour model (context model)
  - User profiling based on previous behaviour (context)
- Representative behavioral patterns
  - Identification of groups with the same behavioral characteristics
  - Try to identify user(s) by using their behavioral patterns only
- Impact on **users' privacy** (ISPs have huge traffic databases available)
- Techniques for finding behavioral characteristics
  - Input data restriction and optimization
  - Processing data (appropriate input information; data mining techniques)
  - Results evaluation → impacts on users' privacy

Introduction

- Customizable services
  - Operate with "user profiles"
  - Reflects user's previous behaviour (based on their context information)
- Context information
  - Descriptive type of information
  - By-product of on-line activity, associated with an individual
  - May reveal some private information
- User behaviour model (context model)
  - User profiling based on previous behaviour (context)
- Representative behavioral patterns
  - Identification of groups with the same behavioral characteristics
  - Try to identify user(s) by using their behavioral patterns only
- Impact on **users' privacy** (ISPs have huge traffic databases available)
- Techniques for finding behavioral characteristics
  - Input data restriction and optimization
  - Processing data (appropriate input information; data mining techniques)
  - Results evaluation → impacts on users' privacy

Introduction

- Customizable services
  - Operate with "user profiles"
  - Reflects user's previous behaviour (based on their context information)
- Context information
  - Descriptive type of information
  - By-product of on-line activity, associated with an individual
  - May reveal some private information
- User behaviour model (context model)
  - User profiling based on previous behaviour (context)
- Representative behavioral patterns
  - Identification of groups with the same behavioral characteristics
  - Try to identify user(s) by using their behavioral patterns only
- Impact on **users' privacy** (ISPs have huge traffic databases available)
- Techniques for finding behavioral characteristics
  - Input data restriction and optimization
  - Processing data (appropriate input information; data mining techniques)
  - Results evaluation → impacts on users' privacy

Introduction

- Customizable services
  - Operate with "user profiles"
  - Reflects user's previous behaviour (based on their context information)
- Context information
  - Descriptive type of information
  - By-product of on-line activity, associated with an individual
  - May reveal some private information
- User behaviour model (context model)
  - User profiling based on previous behaviour (context)
- Representative behavioral patterns
  - Identification of groups with the same behavioral characteristics
  - Try to identify user(s) by using their behavioral patterns only
- Impact on **users' privacy** (ISPs have huge traffic databases available)
- Techniques for finding behavioral characteristics
  - Input data restriction and optimization
  - Processing data (appropriate input information; data mining techniques)
  - Results evaluation → impacts on users' privacy

Content

State-of-the-art

- Context information models
  - Set theory – Context T is described by a set of vectors
  - Directed graph – Something like UML, very comprehensive
  - First-order logic – `Context(<ContextType>,<Subj>,<Rel>,<Obj>)`
- User behaviour models
  - Global mixture model – General model is optimized individually
  - Maximum entropy model – Set of constraints from different sources
- Privacy models
  - Freiburg privacy diamond (FPD) – Mobile environment
  - PATS – Inspired by the FPD but considers all available context information and inner relations
- Models are mainly web oriented
  - Web users' navigational characteristics
  - Input data – web access logs
  - Consider some other type of traffic logs (e.g. SMTP, ftp, ssh, . . .)

State-of-the-art

- Context information models
  - Set theory – Context T is described by a set of vectors
  - Directed graph – Something like UML, very comprehensive
  - First-order logic – `Context(<ContextType>,<Subj>,<Rel>,<Obj>)`
- User behaviour models
  - Global mixture model – General model is optimized individually
  - Maximum entropy model – Set of constraints from different sources
- Privacy models
  - Freiburg privacy diamond (FPD) – Mobile environment
  - PATS – Inspired by the FPD but considers all available context information and inner relations
- Models are mainly web oriented
  - Web users' navigational characteristics
  - Input data – web access logs
  - Consider some other type of traffic logs (e.g. SMTP, ftp, ssh, ...)

State-of-the-art

- Context information models
  - Set theory – Context T is described by a set of vectors
  - Directed graph – Something like UML, very comprehensive
  - First-order logic – `Context(<ContextType>,<Subj>,<Rel>,<Obj>)`
- User behaviour models
  - Global mixture model – General model is optimized individually
  - Maximum entropy model – Set of constraints from different sources
- Privacy models
  - Freiburg privacy diamond (FPD) – Mobile environment
  - PATS – Inspired by the FPD but considers all available context information and inner relations
- Models are mainly web oriented
  - Web users' navigational characteristics
  - Input data – web access logs
  - Consider some other type of traffic logs (e.g. SMTP, ftp, ssh, . . .)

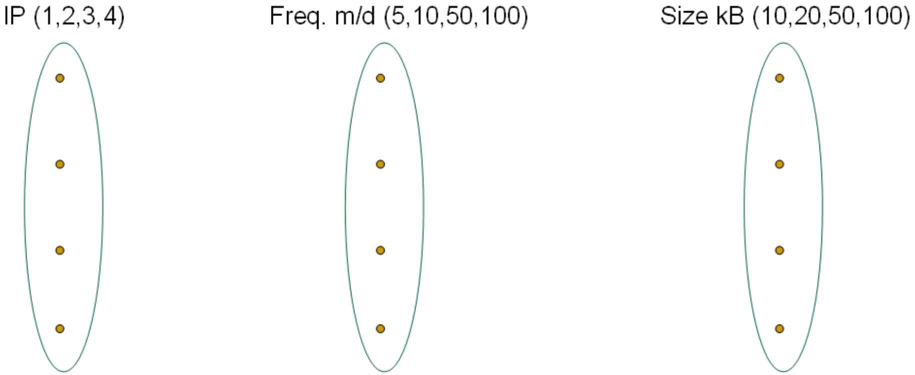
State-of-the-art

- Context information models
  - Set theory – Context T is described by a set of vectors
  - Directed graph – Something like UML, very comprehensive
  - First-order logic – `Context(<ContextType>,<Subj>,<Rel>,<Obj>)`
- User behaviour models
  - Global mixture model – General model is optimized individually
  - Maximum entropy model – Set of constraints from different sources
- Privacy models
  - Freiburg privacy diamond (FPD) – Mobile environment
  - PATS – Inspired by the FPD but considers all available context information and inner relations
- Models are mainly web oriented
  - Web users' navigational characteristics
  - Input data – web access logs
  - Consider some other type of traffic logs (e.g. SMTP, ftp, ssh, ...)

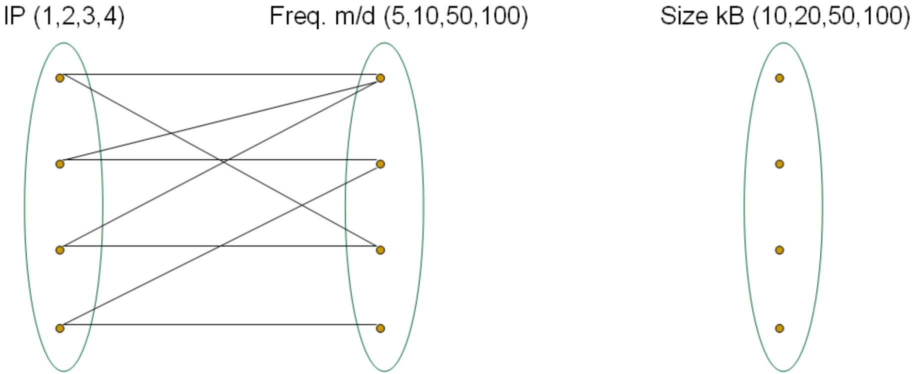
PATS (Privacy Across The Street) model

- Graph represents actual knowledge about a system (context information)
- The goal is to involve all available context information
- Context information is represented as vertices
- Relations between vertices (edges) – weighted with probabilities
- The goal – best (most likely) connection between vertices

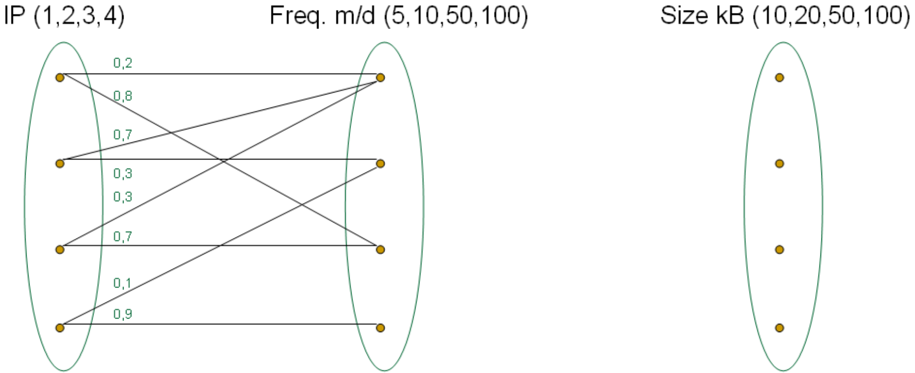
Graph model – an example



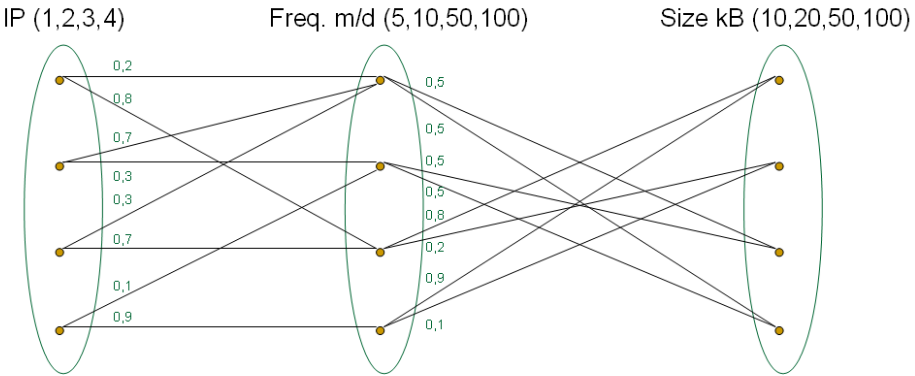
Graph model – an example



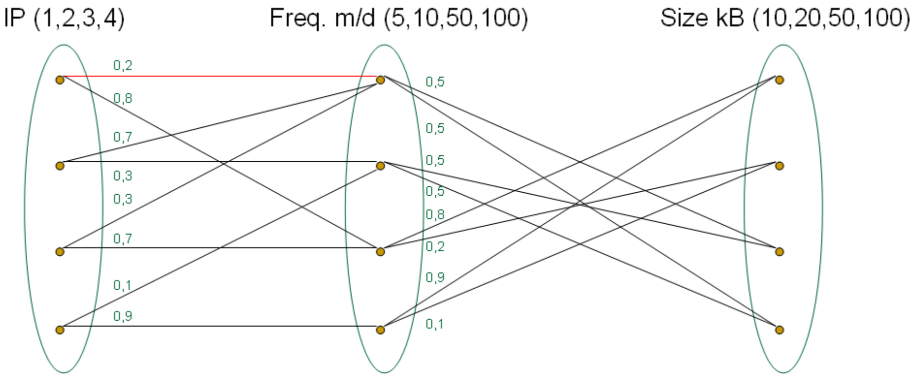
Graph model – an example



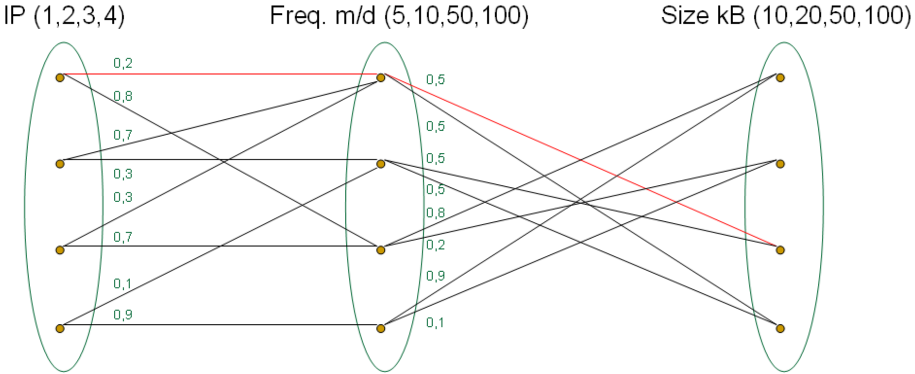
Graph model – an example



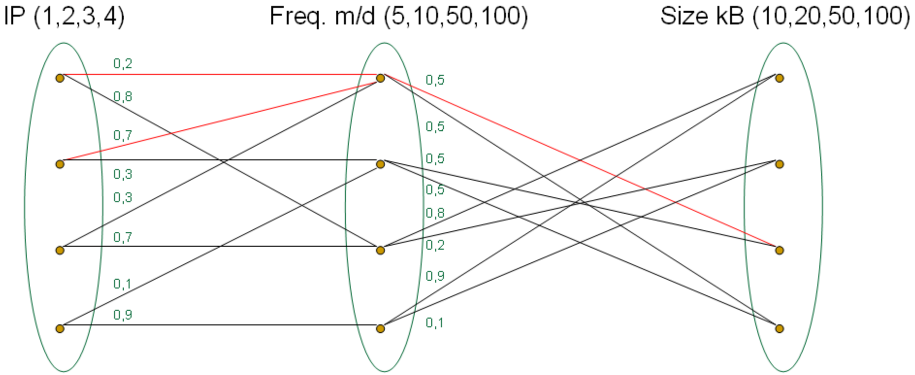
Graph model – an example



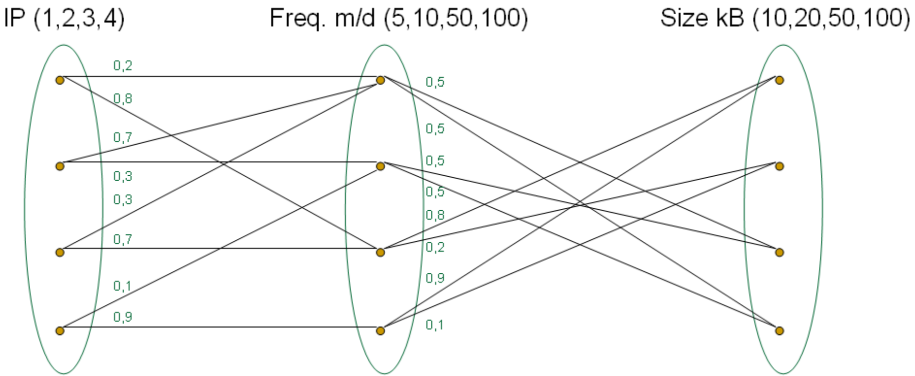
Graph model – an example



Graph model – an example

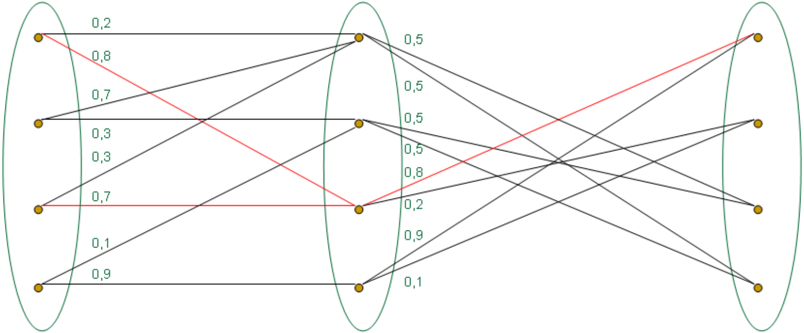


Graph model – an example

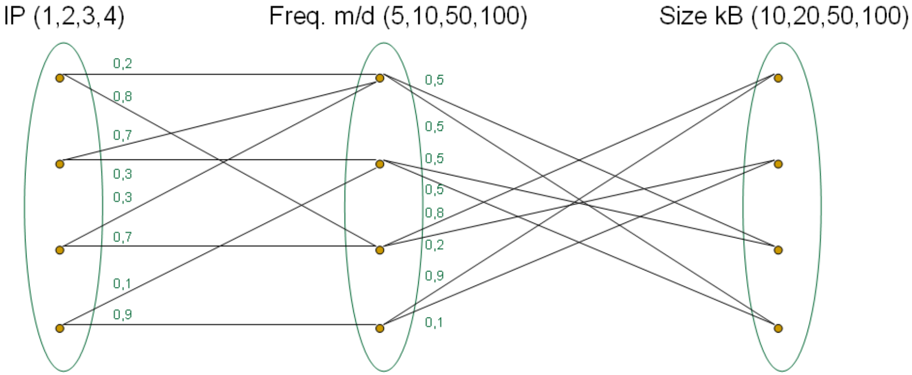


Graph model – an example

IP (1,2,3,4) Freq. m/d (5,10,50,100) Size kB (10,20,50,100)



Graph model – an example



1-5-50-5-2=0,0035

1-50-10-50-3=0,385

Introduction and the story

- AOL released a list of 21 million web search queries on 1. August 06
- Online version `http://www.aolsearchdatabase.com`
- Focused on 658 000 subscribers
- Search queries during a three-month period
- UserIDs were anonymized
- Released on AOL Research site – for academic purposes
- Examples of queries:
  - `find family by social security number`
  - `how to secretly poison your ex`
  - `learning to be single`
- Allows for user profiling – e.g. AOL user 311045 possibly owns a Scion XB automobile in need of new brake pads. User is possibly a Florida resident...
- User 710794 is possibly an overweight golfer, owner of a 1986 Porsche 944 and 1998 Cadillac SLS, and a fan of University of Tennessee Basketball team.

Introduction and the story

- AOL released a list of 21 million web search queries on 1. August 06
- Online version `http://www.aolsearchdatabase.com`
- Focused on 658 000 subscribers
- Search queries during a three-month period
- UserIDs were anonymized
- Released on AOL Research site – for academic purposes
- Examples of queries:
  - `find family by social security number`
  - `how to secretly poison your ex`
  - `learning to be single`
- Allows for user profiling – e.g. AOL user 311045 possibly owns a Scion XB automobile in need of new brake pads. User is possibly a Florida resident...
- User 710794 is possibly an overweight golfer, owner of a 1986 Porsche 944 and 1998 Cadillac SLS, and a fan of University of Tennessee Basketball team.

Identification of a real person

Full identification of a real individual
User No. 4417749 (Thelma Arnold) was identified

Examples of her queries:

- 60 single men
- dog that urinates on everything
- landscapers in Lilburn, Ga
- dogs-related queries

She agreed to discuss her searches with a reporter and was shocked to hear that AOL had saved and published her searches.

How many times did you search your name with Google? :-)

Identification of a real person

Full identification of a real individual
User No. 4417749 (Thelma Arnold) was identified

Examples of her queries:

- 60 single men
- dog that urinates on everything
- landscapers in Lilburn, Ga
- dogs-related queries

She agreed to discuss her searches with a reporter and was shocked to hear that AOL had saved and published her searches.

How many times did you search your name with Google? :-)

Content

Input data – Netflow MU (traffic log)

- Records of communication in MU network (NetFlow)
  - around 180 million records/day
  - source/destination IP; protocol; ports; time; transferred bytes ...
  - current state – over 1 000 000 000 records (one year; many records were dropped)
  - MySQL – problems with speed...

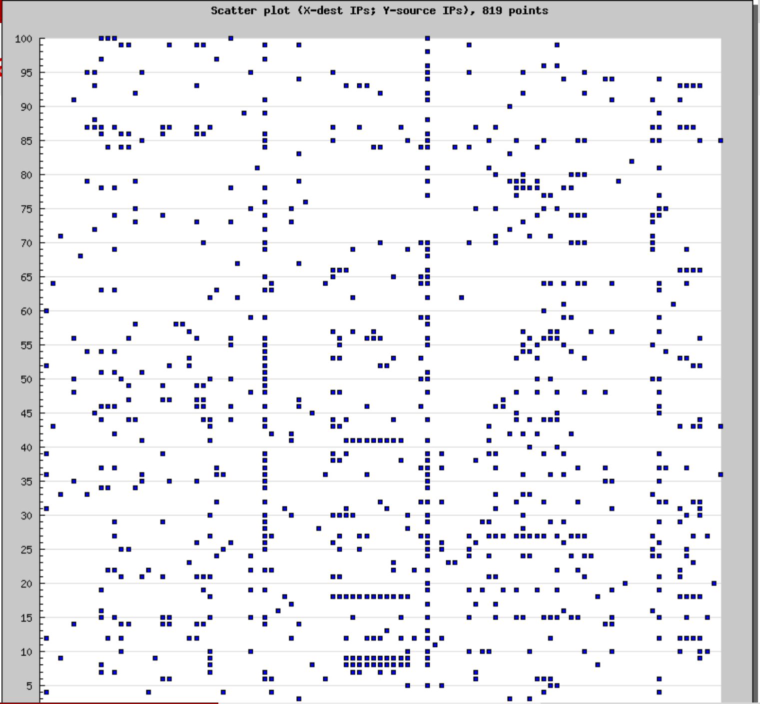
Input data – cont.

- Input restriction – selected part of a network; selected ports (Faculty of informatics and college; port 80, 22)
  - find most frequently visited destination IPs
    - best ratio between source and destination IPs?
    - techniques that help to clear the data
  - for every source IP find the number of hits to a particular destination
- Output is the matrix source vs. destination IPs and hits
  - we have vectors describing "behaviour" of source IPs
  - input data for the clustering process
  - matrix is very sparse :-(
- Approaches to limit the number of context information and entities
  - omit very frequently visited destinations
  - omit commonly visited destinations
  - omit very active source IPs
  - restriction of IP addresses (src/dest) and port
- Input data visualization
  - to visually detect some characteristics

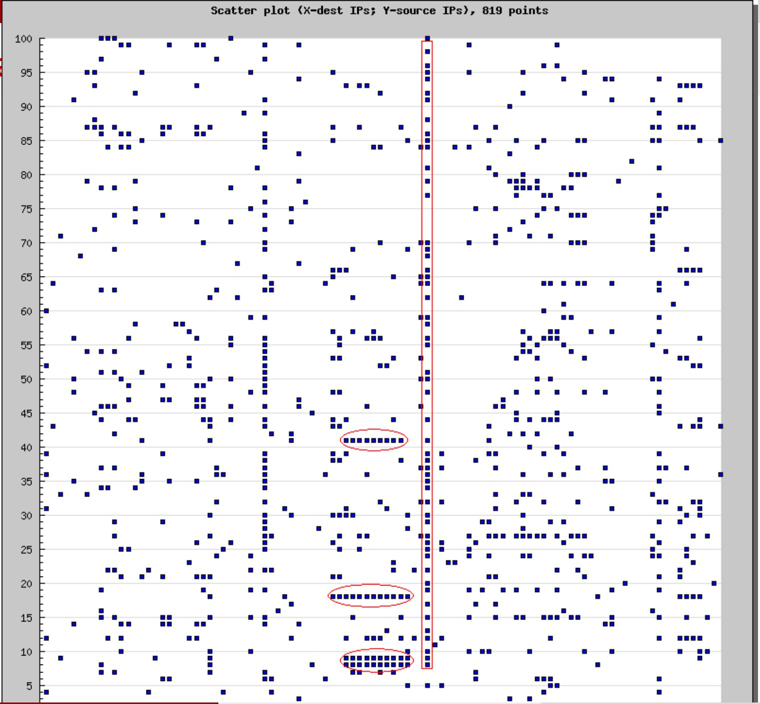
Visualization of input data

- To get an initial view. . .

Visua


Scatter plot (X-dest IPs; Y-source IPs), 819 points
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5

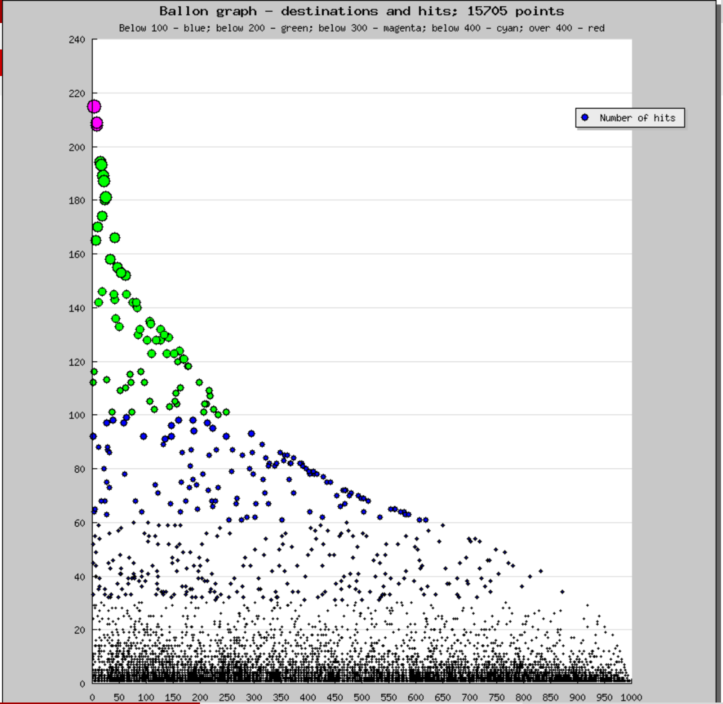
Visua


Scatter plot (X-dest IPs; Y-source IPs), 819 points
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5

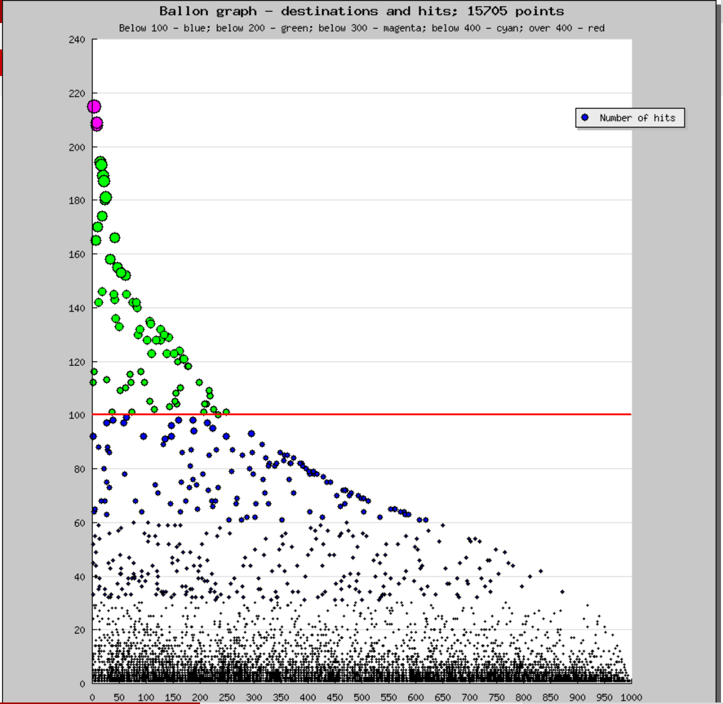
Visualization of input data

- Restrict the number of destination IPs. . .

Visual


Ballon graph – destinations and hits; 15705 points
Below 100 – blue; below 200 – green; below 300 – magenta; below 400 – cyan; over 400 – red
Number of hits
240
220
200
180
160
140
120
100
80
60
40
20
0
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1000

Visual



Ballon graph - destinations and hits; 15705 points
Below 100 - blue; below 200 - green; below 300 - magenta; below 400 - cyan; over 400 - red
Number of hits

Content

Ways to filter input data

How to find relevant source and destination IPs?

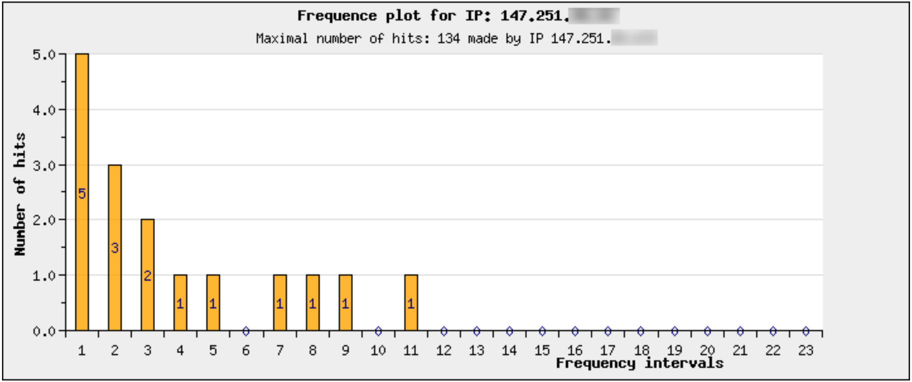
We need more dense matrix for the clustering process

- Destination IPs restrictions
  - accessed only once within a given period
  - accessed by at least a half of sources
  - different levels of entropies – number of unique sources
  - TF-IDF (text mining field), PrefixSpan (sequence based mining)
- Usage-based vs. frequency-based approach
  - usage-based – to optimize destinations
  - frequency-based – to optimize sources
- Visualization of the matrix of vectors
  - scatter plot (usage-based)
  - balloon plot (frequency-based)
- Source IPs restrictions
  - only "active" sources may help in clustering (profiling)
  - behaviour of passive sources is difficult to predict
  - differentiate between different levels of "activity"

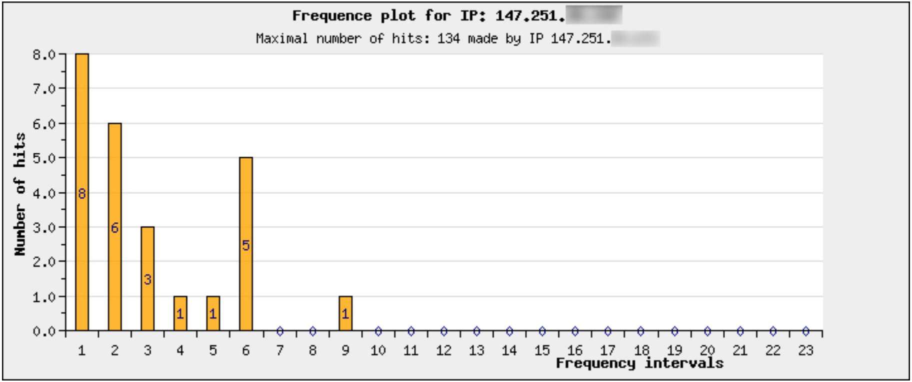
Frequency histograms clustering

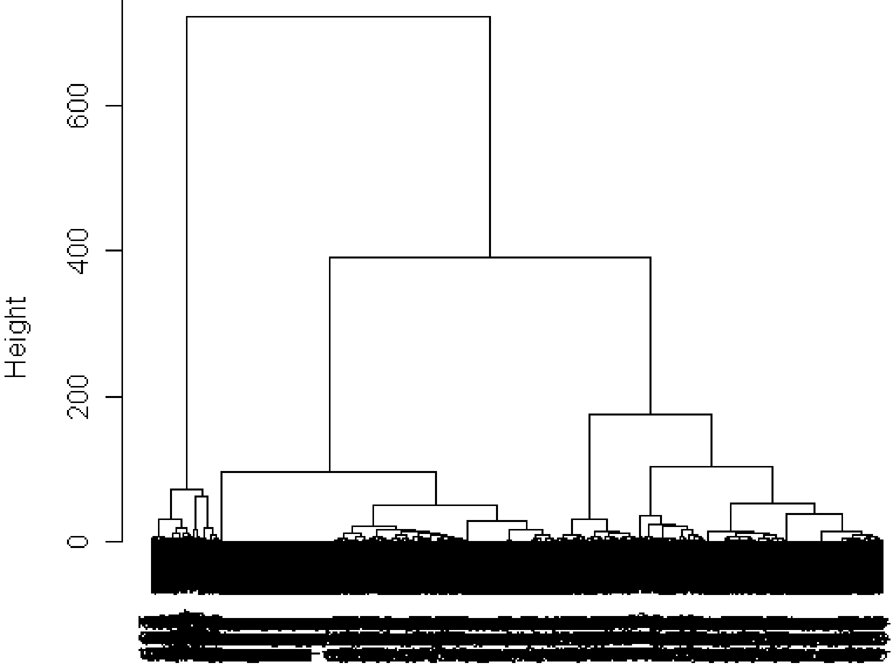
- Frequencies of source IPs activities
  - levels of frequencies and number of accessed destinations
  - 1 to 10 individually and then aggregations of tens
  - most records fall into these individual categories
- Helps to find different levels of activity
- Helps to decrease the matrix dimensions
  - process of clustering is partially automatic
    - ★ find histograms
    - ★ save vectors into arff file
    - ★ use R to perform clustering and cut clusters to sets
  - Ward's clustering method
    - ★ minimizes the 'information loss' associated with each grouping
    - ★ strong tendency to split data in groups of roughly equal size
    - ★ no clusters with only one or a few elements
    - ★ output levels of activity are used as a restriction

Histogram visualization and processing

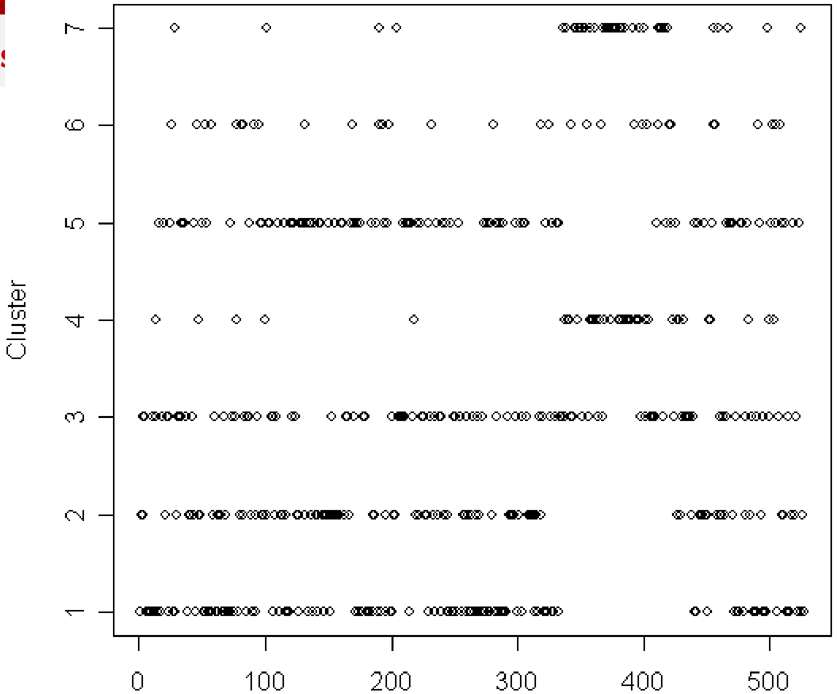


Histogram visualization and processing




Height
0
200
400
600

His



PrefixSpan

- Sequence mining algorithm
- Searching for frequent sequences of destinations
- Sequences can contain gaps (how long?)
- Destinations ordering – IP value
- Input: sequences of destinations for each source
- Output: frequenct sequences w.r.t prefixspan settings
- Frequent sequences can be processed individually
- . . . to find corresponding sources
- Sources can be analyzed with more data
- Problems with proxies and very active sources

```
./prefixspan -m 2 -M 5 <sequences.txt >output.txt
-m NUM:   set minimum support
-M NUM:   set minimum pattern length
-L NUM:   set maximum pattern length
-a:       print ALL patterns (default: print longest pattern)
```

Content

Similarity computation – cosine similarity

- Data from two time periods (e.g. months)
- First dataset – apply some restrictions → 1st temp. table
- Second dataset – apply the same restriction → 2nd temp. table
- Different types of restrictions and their influence
- IDF values based on the first table – highly dependent information
- Synchronize temp. tables – vectors of the same dimensions (set of destinations)
- Cosine similarity measure (of two behavioural vectors A, B)
  - $cosim(\varphi) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$
  - 1 – completely related; 0 – completely unrelated
  - For every vector from the 1st table → list of candidates
    - ⋆ $A : \ldots, sim\_value_{(A,B)}(B, d_{comm}), \ldots$

Similarity computation – cosine similarity

- Data from two time periods (e.g. months)
- First dataset – apply some restrictions → 1st temp. table
- Second dataset – apply the same restriction → 2nd temp. table
- Different types of restrictions and their influence
- IDF values based on the first table – highly dependent information
- Synchronize temp. tables – vectors of the same dimensions (set of destinations)
- Cosine similarity measure (of two behavioural vectors A, B)
  - $cosim(\varphi) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$
  - 1 – completely related; 0 – completely unrelated
  - For every vector from the 1st table → list of candidates
    - ⋆ $A : \ldots, sim\_value_{(A,B)}(B, d_{comm}), \ldots$

An example

A	1(A,1); 1(B,1); 1(O,1); 0.164399(E,1)
B	1(A,1); 1(B,1); 1(O,1); 0.164399(E,1)
D	0.999635(M,1); 0.997976(D,2); 0.0270172(J,1)
E	0.999168(E,2); 0.124035(A,1); 0.124035(B,1); 0.124035(O,1)
J	1(J,1); 0.0905358(D,1)

A	1(A,1); 1(B,1); 1(O,1); 0.0763637(E,1)
B	1(A,1); 1(B,1); 1(O,1); 0.0763637(E,1)
D	0.999806(M,1); 0.998918(D,2); 0.0197195(J,1)
E	0.999818(E,2); 0.057345(A,1); 0.057345(B,1); 0.057345(O,1)
J	1(J,1); 0.0661965(D,1)

	1	2	3	4	5	6	7	8	9
A	0	0	0	0	0	0	0	0	1853
B	0	0	0	0	0	0	0	0	297
D	0	0	37	0	0	0	1	0	0
E	0	0	0	0	32	0	0	0	4
J	0	0	0	0	0	0	17	0	0

	1	2	3	4	5	6	7	8	9
A	0	0	0	0	0	0	0	0	1487
B	0	0	0	0	0	0	0	0	244
E	0	0	0	0	12	0	0	0	2
J	0	0	0	0	0	0	12	0	0
D	0	0	11	0	0	0	1	0	0
M	0	0	5	0	0	0	0	0	0
O	0	0	0	0	0	0	0	0	3

An example

A	1(A,1); 1(B,1); 1(O,1); 0.164399(E,1)
B	1(A,1); 1(B,1); 1(O,1); 0.164399(E,1)
D	0.999635(M,1); 0.997976(D,2); 0.0270172(J,1)
E	0.999168(E,2); 0.124035(A,1); 0.124035(B,1); 0.124035(O,1)
J	1(J,1); 0.0905358(D,1)

A	1(A,1); 1(B,1); 1(O,1); 0.0763637(E,1)
B	1(A,1); 1(B,1); 1(O,1); 0.0763637(E,1)
D	0.999806(M,1); 0.998918(D,2); 0.0197195(J,1)
E	0.999818(E,2); 0.057345(A,1); 0.057345(B,1); 0.057345(O,1)
J	1(J,1); 0.0661965(D,1)

	1	2	3	4	5	6	7	8	9
A	0	0	0	0	0	0	0	0	1853
B	0	0	0	0	0	0	0	0	297
D	0	0	37	0	0	0	1	0	0
E	0	0	0	0	32	0	0	0	4
J	0	0	0	0	0	0	17	0	0

	1	2	3	4	5	6	7	8	9
A	0	0	0	0	0	0	0	0	1487
B	0	0	0	0	0	0	0	0	244
E	0	0	0	0	12	0	0	0	2
J	0	0	0	0	0	0	12	0	0
D	0	0	11	0	0	0	1	0	0
M	0	0	5	0	0	0	0	0	0
O	0	0	0	0	0	0	0	0	3

An example

A	1(A,1); 1(B,1); 1(O,1); 0.164399(E,1)
B	1(A,1); 1(B,1); 1(O,1); 0.164399(E,1)
D	0.999635(M,1); 0.997976(D,2); 0.0270172(J,1)
E	0.999168(E,2); 0.124035(A,1); 0.124035(B,1); 0.124035(O,1)
J	1(J,1); 0.0905358(D,1)

A	1(A,1); 1(B,1); 1(O,1); 0.0763637(E,1)
B	1(A,1); 1(B,1); 1(O,1); 0.0763637(E,1)
D	0.999806(M,1); 0.998918(D,2); 0.0197195(J,1)
E	0.999818(E,2); 0.057345(A,1); 0.057345(B,1); 0.057345(O,1)
J	1(J,1); 0.0661965(D,1)

	1	2	3	4	5	6	7	8	9
A	0	0	0	0	0	0	0	0	1853
B	0	0	0	0	0	0	0	0	297
D	0	0	37	0	0	0	1	0	0
E	0	0	0	0	32	0	0	0	4
J	0	0	0	0	0	0	17	0	0

	1	2	3	4	5	6	7	8	9
A	0	0	0	0	0	0	0	0	1487
B	0	0	0	0	0	0	0	0	244
E	0	0	0	0	12	0	0	0	2
J	0	0	0	0	0	0	12	0	0
D	0	0	11	0	0	0	1	0	0
M	0	0	5	0	0	0	0	0	0
O	0	0	0	0	0	0	0	0	3

Sim. measure – proposed improvements

- General idea – strengthen rare attributes
  - Knowledge of a certain rare attribute vs. common attribute
  - Same idea used later by Narayanan and Shmatikov

- TF-IDF (Term Frequency - Inverse Document Frequency)
  - IDF – how important a destination is to a set of source IPs
  - $weight(i,j) = tf_{i,j} \cdot log_2(n/df_i),\ \text{if } tf_{i,j} \geq 1$
  - Highly dependent on current structure of input data
  - Additional context information for a given “environment”
  - Vector of relevance (same size as behavioural vectors)
  - Multiplied with all behavioural vectors (prior *cosim*)

Sim. measure – proposed improvements

- d_{comm} values – number of common attributes (destination IPs)
  - $\nearrow$ num. of common destinations $\Rightarrow$ $\nearrow$ similarity index
  - Re-computed after the main similarity searching procedure
  - HTTP traffic – average number of common attributes – 3.3
  - $d_{(A,B)} = d_{comm}/d_{max}$
  - $sim\_value_{(A,B)} = \frac{cos(A,B)+d_{(A,B)}}{2}$
- Comparison with Narayanan and Shmatikov
  - *Robust De-anonymization of Large Sparse Datasets* (IEEE, 2008)
  - Knowledge of 3-8 shared attributes for re-identification
  - Same approach for strengthening rare attributes
  - More dense data (movie rating DB)
  - Our experiments: SSH – 1.5; HTTPS – 6; HTTP – 3.3

Sim. measure – proposed improvements

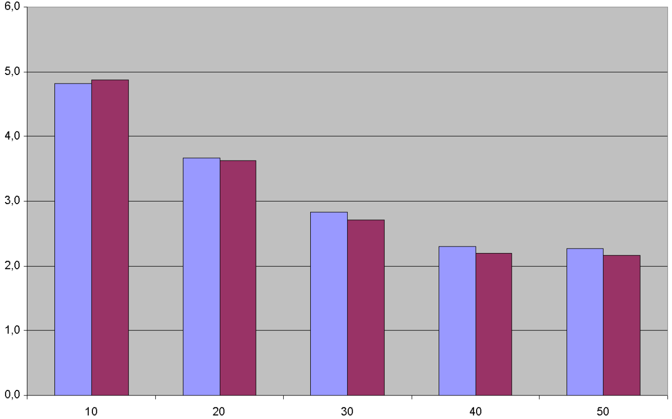


Figure: HTTP traffic – Training and testing sets (average number of visited destination IP addresses)

Similarity measure – evaluation

- Evaluation of our two proposed improvements (IDF and d_{comm})
- Different initial conditions and their impact

- Two proposals for evaluation:
  - Comparison with the "ideal" model
    - ★ We know the correct answer from the original data
    - ★ Distance between the correct answer and the output of the similarity measure
    - ★ We can observe the influence of IDF values
  - Evaluation based on three criteria
    - ★ "Correct" candidate is the first on the list of candidates
    - ★ "Correct" candidate is in the list of candidates (but not the first)
    - ★ "Correct" is not in the testing set
- Evaluation of profiles' persistence
  - Always fresh profiles (e.g., neighboring months)
  - Old profiles (e.g., created in January)

Similarity measure – comparison with ideal model

- Normalize the set of similar IPs – sum equals 1
- $|1 - sim\_index|$ – correct decision (we "know" which one is correct)
- $|0 - sim\_index|$ – bad decision
- Sum of these for every source IP – "amount of error"

A	1(A,1); 1(B,1); 1(O,1); 0.164399(E,1)
A	0.316015(A,1); 0.316015(B,1); 0.316015(O,1); 0.051953(E,1)
A	1(A,1); 0(B,1); 0(O,1); 0(E,1)
A	0.683985(A,1); 0.316015(B,1); 0.316015(O,1); 0.051953(E,1)

- Error rate – 1.367968 (boundaries – $0 \rightarrow 2$)

The influence of the IDF values

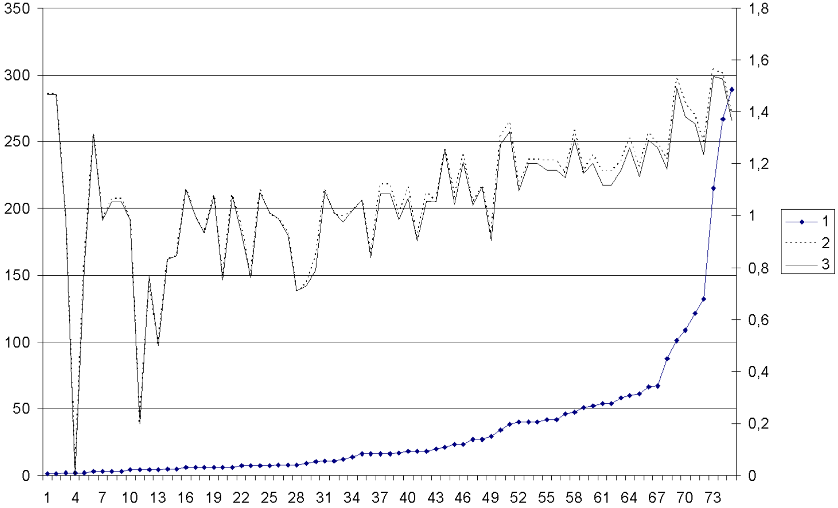


Figure: Influence of the IDF – HTTP traffic.

Evaluation based on the three criteria (HTTP)

restr.	crit.	IDF + $d_{(A,B)}$	IDF	$\cos(A, B)$
20	"Correct" – 1^{st} place	20%	10%	14%
20	"Correct" – in the list	17%	27%	22%
⋮	⋮	⋮	⋮	⋮

Table: HTTP traffic – first and second criteria (shortened)

- Number of common attributes for a 100% re-identification – 3.3
- Third criteria (candidate not in the testing set) – 61.5%
- Average distance from the first candidate (second crit.) – 0.12
- IDF + $d_{(A,B)}$ move the correct candidates to the beginning in the list of candidates

Stability of user profiles

- How long is a user profile “fresh”?
- . . . and can be used for re-identification
- Two experiments:
  1. Training and testing sets are neighbouring months
  2. First month (only) of a year used as a training set
- Results (decrease caused by old profiles):
  1. SSH traffic – 9.89 %
  2. HTTPS traffic – 5.15 %
  3. HTTP traffic – 13.36 %

Stability of user profiles

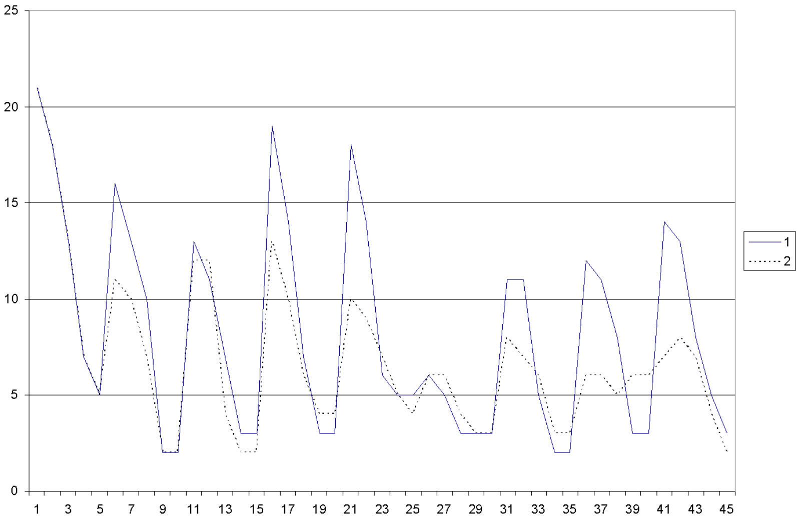


Figure: Stability of profiles based on SSH traffic.

Stability of user profiles

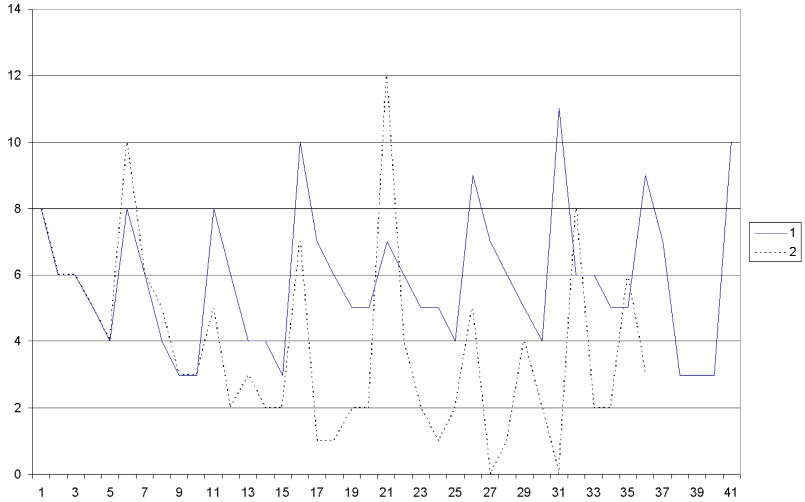


Figure: Stability of profiles based on HTTPS traffic.

Stability of user profiles

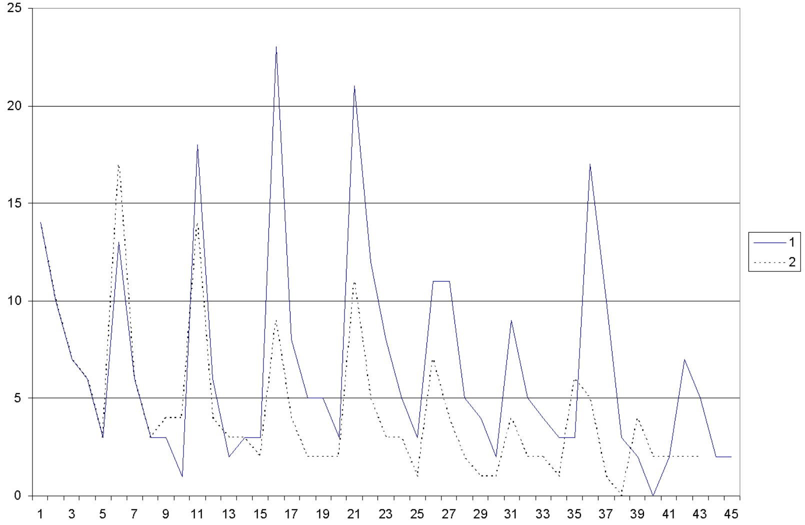


Figure: Stability of profiles based on HTTP traffic.

Conclusions and ideas for future research

- Main contribution of the project
  - PATS model for context information analysis
  - Experiments towards re-identification with real data
    - ⋆ Two proposed improvements of the cosine similarity measure
    - ⋆ IDF and d_{comm} values
  - Evaluation of the similarity searching procedure
    - ⋆ IDF and d_{comm} values provide better results
    - ⋆ Evaluation of the measure for SSH, HTTPS and HTTP protocols
    - ⋆ Overall re-identification rates – 58.61%, 19.67%, 19.33%
- Ideas for the future research:
  - Further evaluations; stability of user profiles
  - Another approach of building behavioural vectors – progressively in time
  - Different input data

Questions?

Thanks for your attention!

LaTeX