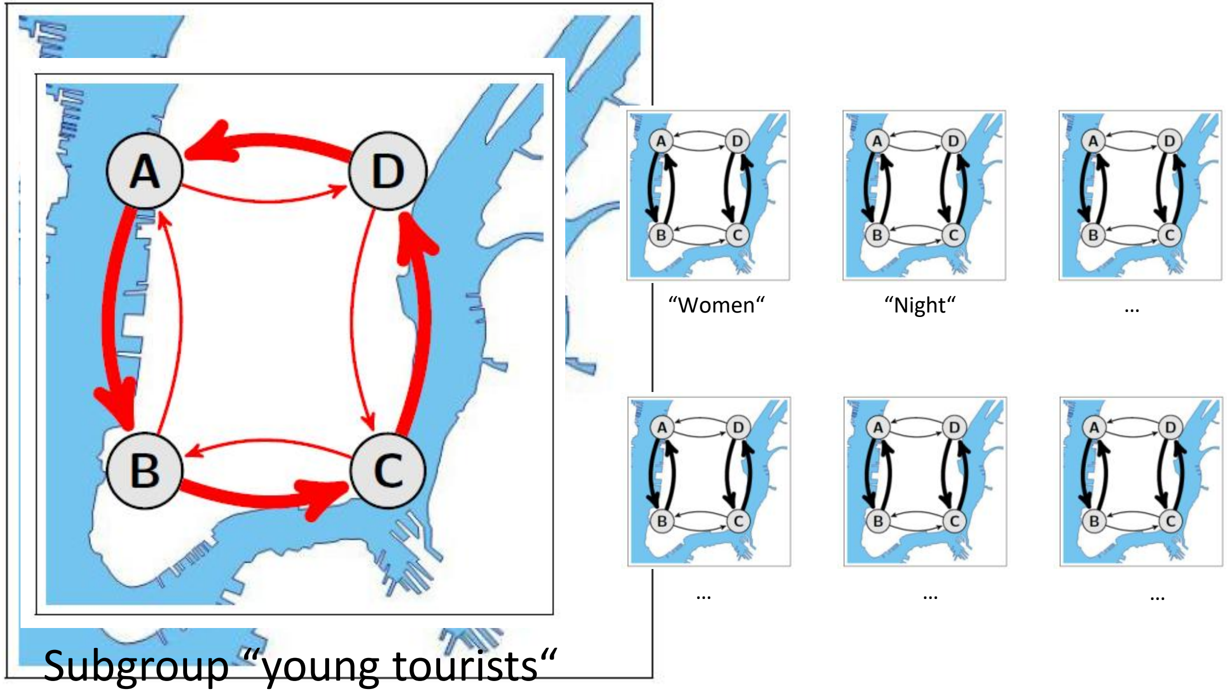
Mining Subgroups with Exceptional Transition Behavior

Florian Lemmerich[1], Martin Becker[2], Philipp Singer[1]
Denis Helic[3], Andreas Hotho[2,4], Markus Strohmaier[1]

[1] GESIS – Leibniz Institute for the Social Sciences, Cologne & University of Koblenz-Landau
[2] University of Würzburg
[3] TU Graz
[4] L3S Research Center, Hannover

San Francisco - KDD - 17th of August 2016

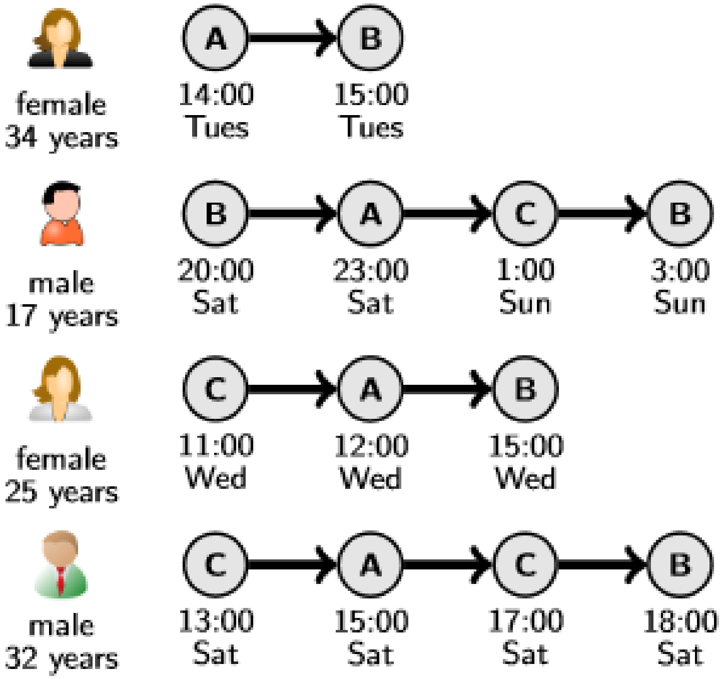
Transition Behavior



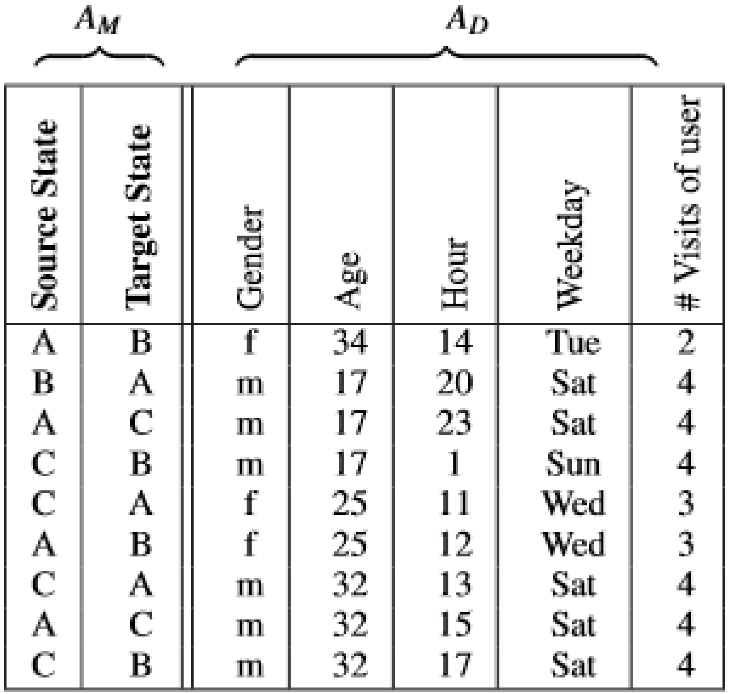
Overall dataset

Subgroup "young tourists"

Data Preparation



(a) Sequence data with background knowledge



A_M		A_D				
Source State	**Target State**	Gender	Age	Hour	Weekday	# Visits of user
A	B	f	34	14	Tue	2
B	A	m	17	20	Sat	4
A	C	m	17	23	Sat	4
C	B	m	17	1	Sun	4
C	A	f	25	11	Wed	3
A	B	f	25	12	Wed	3
C	A	m	32	13	Sat	4
A	C	m	32	15	Sat	4
C	B	m	32	17	Sat	4

(b) Transition dataset

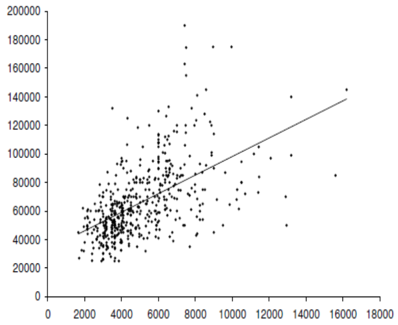
Exceptional Model Mining

Task

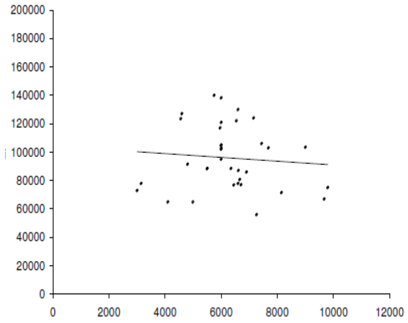
Find ***DESCRIPTIONS*** of subsets of the data that imply ***EXCEPTIONAL*** (=significantly different) ***PARAMETERS*** with respect to a certain ***MODEL CLASS***.

Example

“In the overall data,
there is a strong correlation between
duration *and* ***distance_travelled***
This is not the case for the subgroup
age > 60 ∧ country = US“

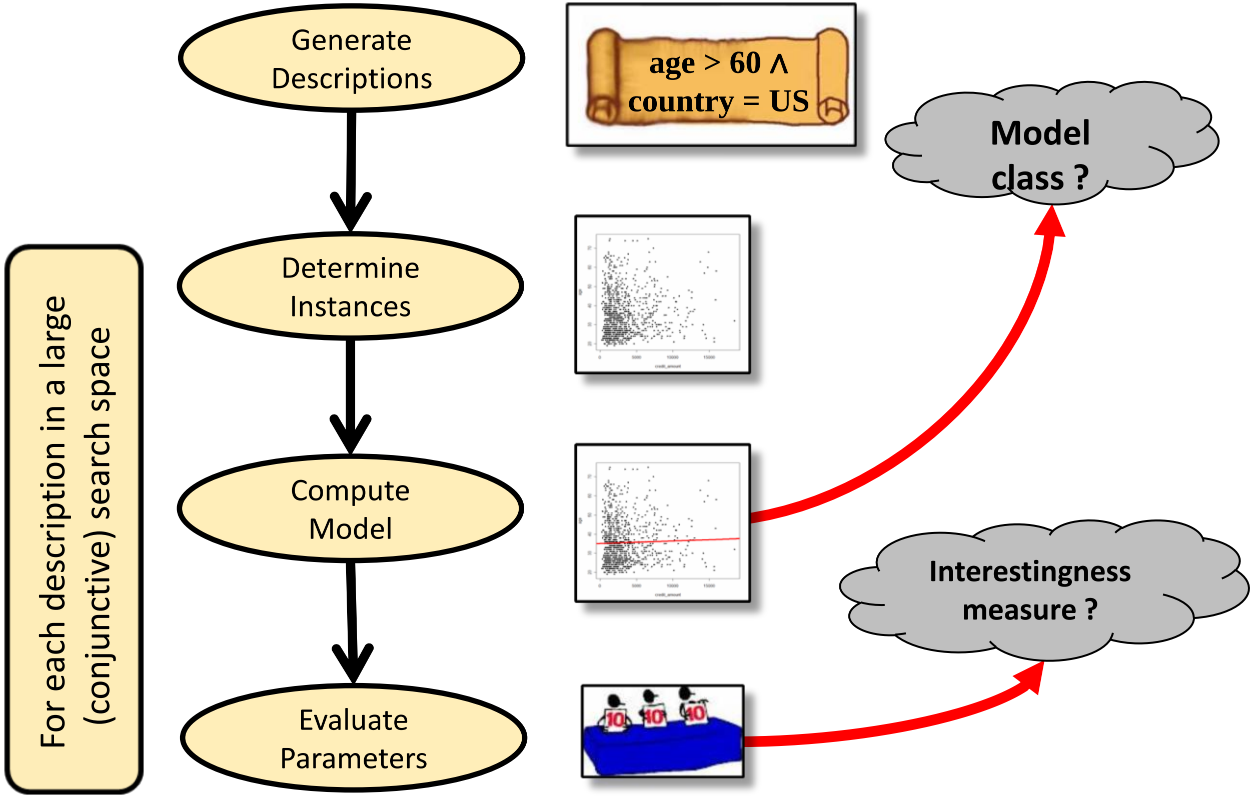


Overall Dataset



Subgroup
age > 60 ∧ country = US

Exceptional Model Mining: Approach



For each description in a large (conjunctive) search space

Generate Descriptions

age > 60 ∧ country = US

Determine Instances

Compute Model

Evaluate Parameters

Model class ?

Interestingness measure ?

Studied model classes:

- *Correlation coefficient*
- *Linear regression*
- *Decision rules (classifier)*
- *Bayesian networks*
- *…*
- ***No models for sequential data***

We need another model class!

New model class requires

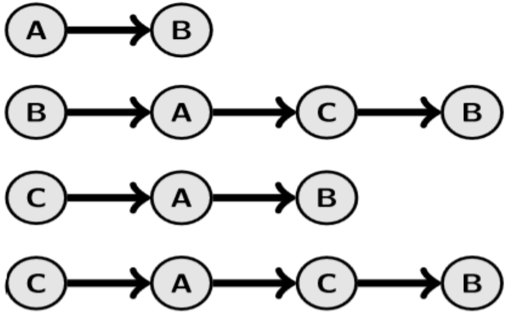
new interestingness measure!

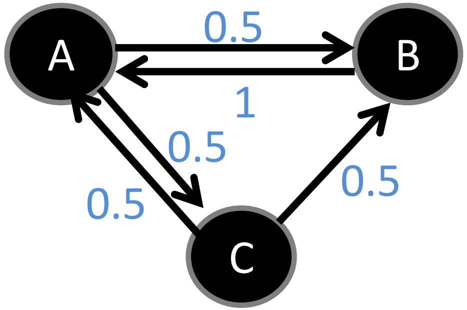
First-order Markov Chains

- Model for sequential data (sequences of states)
- Memoryless process:

  Probability of the next state depends only on the current state
- Well established and frequently used in many areas:

  Human mobility & navigation, economics, metereology, …





$$\begin{pmatrix} 0 & \boxed{0.5} & 0.5 \\ 1 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \end{pmatrix}$$

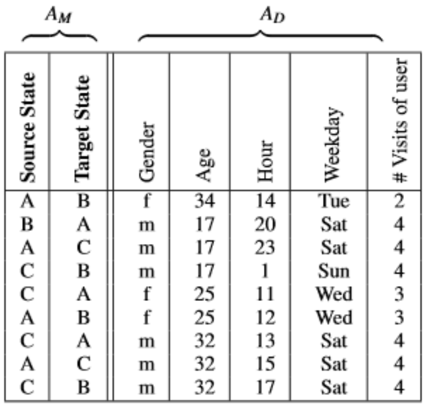
Interestingness Measure: General

- Create “score“ for each candidate subgroup
- Reflects how interesting/exceptional the transitions for a subgroup is
- Search algorithms: Return the *k* subgroups with the best scores

- Interestingness measure are subjective, but…
- … should be able to distinguish influence factors from random noise (in artificial data)

Interestingness Measure: Distance Measure

- Compute transition count matrix for *dataset*, transition probability matrix
- Compute transition count matrix for *subgroup*, transition probability matrix
- Compare rowwise (Manhattan-distance, KL-divergence, Hellinger distance)
- Weight with #transition in this row (subgroup)



A_M		A_D				
Source State	Target State	Gender	Age	Hour	Weekday	# Visits of user
A	B	f	34	14	Tue	2
B	A	m	17	20	Sat	4
A	C	m	17	23	Sat	4
C	B	m	17	1	Sun	4
C	A	f	25	11	Wed	3
A	B	f	25	12	Wed	3
C	A	m	32	13	Sat	4
A	C	m	32	15	Sat	4
C	B	m	32	17	Sat	4

Dataset

$$\begin{pmatrix} 0 & 2 & 2 \\ 1 & 0 & 0 \\ 2 & 2 & 0 \end{pmatrix} \longrightarrow \begin{pmatrix} 0 & 0.5 & 0.5 \\ 1 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \end{pmatrix}$$

Subgroup (gender = f)

$$\begin{pmatrix} 0 & 2 & 0 \\ - & - & - \\ 1 & 0 & 0 \end{pmatrix} \longrightarrow \begin{pmatrix} 0 & 1 & 0 \\ - & - & - \\ 1 & 0 & 0 \end{pmatrix}$$

difference in that row

# transitions in that row

$$\begin{pmatrix} 1 \\ - \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} = 3$$

Interestingness Measure: Distance M

- Compute transition count matrix for *dataset* atrix
- Compute transition count matrix for *su* atrix
- Compare rowwise (Manhattan-di e)
- Weight with #transition i

Problem:

Distance is heavily influenced by subgroup size

A_M		A_D				
Source State	Target State	Gender	Age	Hour	Weekday	# Visits of
A	B	f	34	14	Tue	2
B	A	m	17	20	Sat	4
A	C	m	17	23	Sat	4
C	B	m	17	1	Sun	4
C	A	f	25	11	Wed	3
A	B	f	25	12	Wed	3
C	A	m	32	13	Sat	4
A	C	m	32	15	Sat	4
C	B	m	32	17	Sat	4

nce
that row

# transitions
in that row

$$0.5\Big)$$

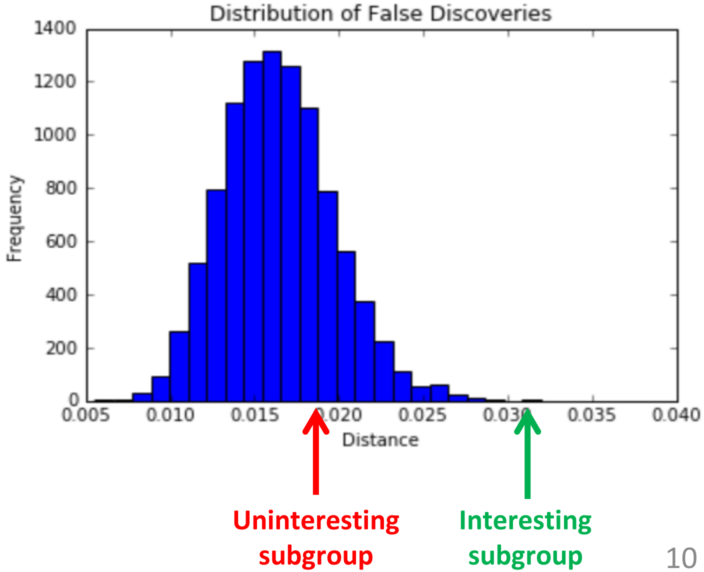
$$\begin{pmatrix} 1 \\ - \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} = 3$$

Sub

$$0\Big) \rightarrow \begin{pmatrix} 0 & 1 & 0 \\ - & - & - \\ 1 & 0 & 0 \end{pmatrix}$$

Interestingness Measure: Random Samples

- Draw *stratified random samples* R_1, … R_r (same size as subgroup)
- For each random sample compute the distance
- Build a *distribution of false discoveries* from the sample distances
- Compute z-score of the distance of the subgroup:
- z-score (g, D) = $\dfrac{\omega_{tv}(g,D) - \mu(\omega_{tv}(R_1,D),\ldots,\omega_{tv}(R_r,D))}{\sigma(\omega_{tv}(R_1,D),\ldots,\omega_{tv}(R_r,D)) + \varepsilon}$



Uninteresting subgroup

Interesting subgroup

Interestingness Measure: Additional Issues

- How many random samples?
  - Speed ⇔ Accuracy
  - Difficult to say in general
  - Can estimate precision of the result with bootstrapping procedure

- How to check significance of findings?
  - If distribution of false discoveries approx. normal: Use z-score directly
  - Otherwise: Draw more samples, compute empirical p-value
  - Always apply Bonferroni-adjustment for multiple comparisons

Interestingness Measure: Summary

*“How different is the distance between
the parameter matrix of the subgroup and the matrix of the dataset
compared to the respective distances of stratified random samples?“*

Subgroup Search

- Any search algorithm for EMM can be employed:
  - Depth-First-Search
  - Best-First-Search
  - Beam-Search
  - …
- Evaluation of one candidate subgroup: $O\left(\overset{\text{\# random samples}}{r} * \left(\overset{\text{\# transitions in the dataset}}{N} + \overset{\text{\# states in the dataset}}{S^2}\right)\right)$

Investigate Hypotheses

- Recently, **user-defined hypotheses** on state-transitions have been studied
  Example for mobility data:
  - “People navigate to a state, which is nearby and contain a Point-of-Interest”
  - “People all move from A to B, from B to C, and from C to A”
- Hypothesis is formalized as a matrix of transition probabilities

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

“People always move from A to B,
from B to C,
and from C to A”

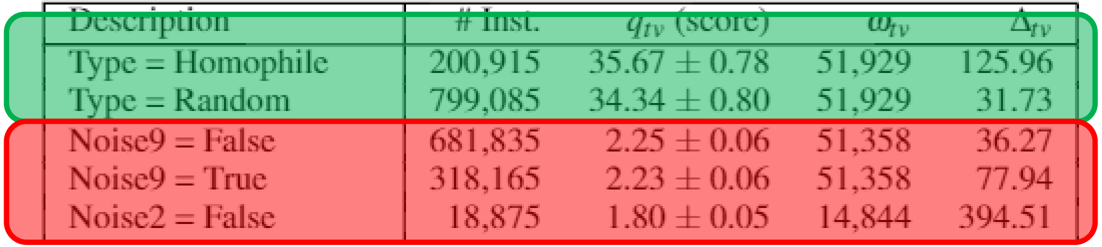
- Can use same approach to find subgroups that **match/contradict** this hypothesis:
  - Replace dataset transition matrix with hypothesis
  - High score: Subgroup deviates from the hypothesis exceptionally strong
  - Low score: Subgroup matches the hypothesis exceptionally well

Evaluation: Synthetic Data, Random Walker example

- Network of colored nodes
- Generate transitions with Random Walkers
  - Source/target of the transition (model attributes)
  - Type of the Walker
  - Randomly assigned noise attributes
- Each random walker in the network with 2 types:
  - “Random“: Transition to all neighbors with same probability
  - “Homophile“: Transition with higher probability to the same color
- Goal:
  - Identify the walker type as the attribute that influences transitions
  - Which subgroups match/predict a “homophile” hypothesis

Results: Random Walker

Subgroups deviating from the dataset:

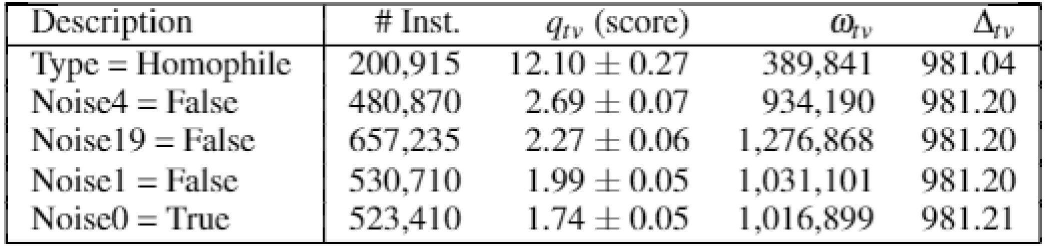


Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
Type = Homophile	200,915	35.67 ± 0.78	51,929	125.96
Type = Random	799,085	34.34 ± 0.80	51,929	31.73
Noise9 = False	681,835	2.25 ± 0.06	51,358	36.27
Noise9 = True	318,165	2.23 ± 0.06	51,358	77.94
Noise2 = False	18,875	1.80 ± 0.05	14,844	394.51

Strongly significant even for weak signals!

Non-significant for noise

Subgroups that are most “homophile“:



Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
Type = Homophile	200,915	12.10 ± 0.27	389,841	981.04
Noise4 = False	480,870	2.69 ± 0.07	934,190	981.20
Noise19 = False	657,235	2.27 ± 0.06	1,276,868	981.20
Noise1 = False	530,710	1.99 ± 0.05	1,031,101	981.20
Noise0 = True	523,410	1.74 ± 0.05	1,016,899	981.21

Application Example: Flickr

- Sequences of user locations derived from uploaded pictures (in Manhattan)
- Locations mapped to tracts (areal units) for discrete state space

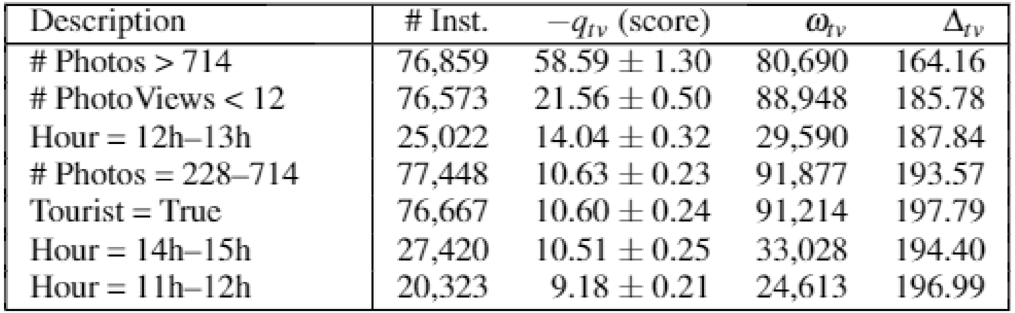
- Attributes:
  - # Photos
  - # Views of picture
  - Months, weekday, hour of the day
  - Tourist/non-tourist
  - Country of origin of the user

- Additional hypothesis: PROXIMATE POI

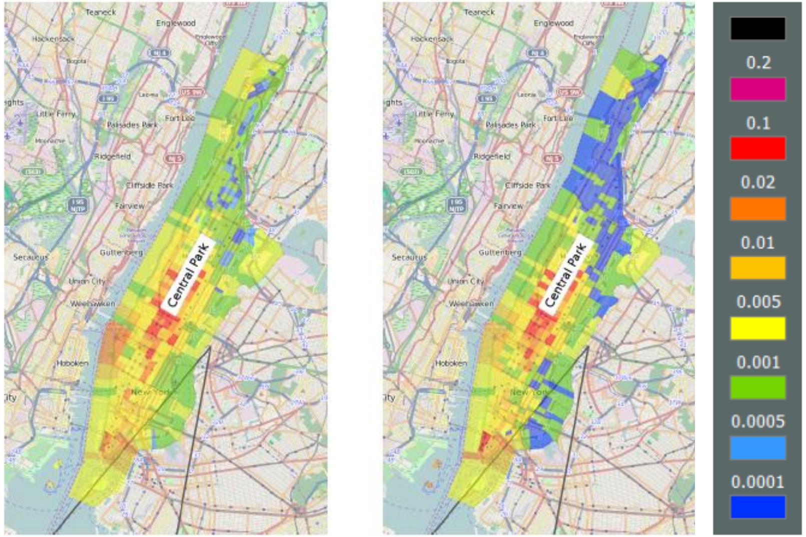
  “People navigate to a state, which is nearby and contain a Point-of-Interest

Results: Flickr

Subgroup matching the PROXIMATE-POI hypothesis:



Description	# Inst.	$-q_{tv}$ (score)	ω_{tv}	Δ_{tv}
# Photos > 714	76,859	58.59 ± 1.30	80,690	164.16
# PhotoViews < 12	76,573	21.56 ± 0.50	88,948	185.78
Hour = 12h–13h	25,022	14.04 ± 0.32	29,590	187.84
# Photos = 228–714	77,448	10.63 ± 0.23	91,877	193.57
Tourist = True	76,667	10.60 ± 0.24	91,214	197.79
Hour = 14h–15h	27,420	10.51 ± 0.25	33,028	194.40
Hour = 11h–12h	20,323	9.18 ± 0.21	24,613	196.99



(a) All transitions

(b) Transitions of tourists

Conclusion

- New approach for mining subgroup with exceptional transition behavior

- Apply Exceptional Model Mining with first-order Markov chain models
- Interestingness measure for this model class
- Can also investigate user-defined hypotheses

- Evaluation:
  - Tested with synthetic data
  - Demonstrated usage in real-world scenarios

Mining Subgroups with Exceptional Transition Behavior

Thank you!

Florian Lemmerich[1], Martin Becker[2], Philipp Singer[1]

Denis Helic[3], Andreas Hotho[2,4], Markus Strohmaier[1]

[1] GESIS – Leibniz Institute for the Social Sciences, Cologne & University of Koblenz-Landau

[2] University of Würzburg

[3] TU Graz

[4] L3S Research Center, Hannover

San Francisco - KDD - 17th of August 2016