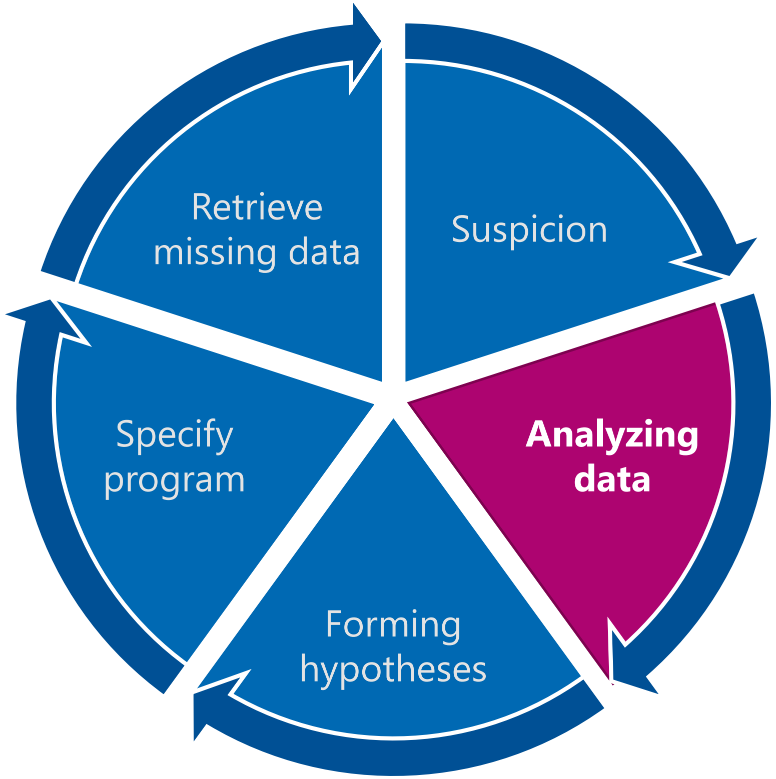
HOCHSCHULE MITTWEIDA
University of Applied Sciences

Tutorial Datasys 2020

Text Mining as a Tool in Repressive and Preventive Investigation Process

Michael Spranger

Criminalistic Cycle



Current
mainly **manual**

Specialized algorithms:

- Text mining
- Information extraction
- Knowledge representation

Forensic
(textual) information management

Real World Case

Prosecutor General's Office Hamburg

Investigation for support of a terrorist group

29,823 messages / 351 chats

9,735 messages / 39 Chats

323 messages / 293 chats

13,665 messages / 41 chats

27,578 messages / 640 chats

5,093 messages / 381 chats

132,640 messages / 1432 chats

7,986 messages / 794 chats

weeks

total: **226,843** messages in **3,971** chats to analyze

minutes

HOCHSCHULE MITTWEIDA

Is text mining well researched in this domain?

- well-formed **english** texts
- closed, **limited** domains

TEXT MINING

- retrieval/filtering of relevant documents
- obtaining (forensic) information
- visualizing (forensic) information

- non-english forensic texts
- **interdisciplinary** domains

Challenges

- rich of slang
- little context
- socio-economically shaped
- heterogeneous
- hidden semantics
- language-economically eroded

+

KNOWLEDGE

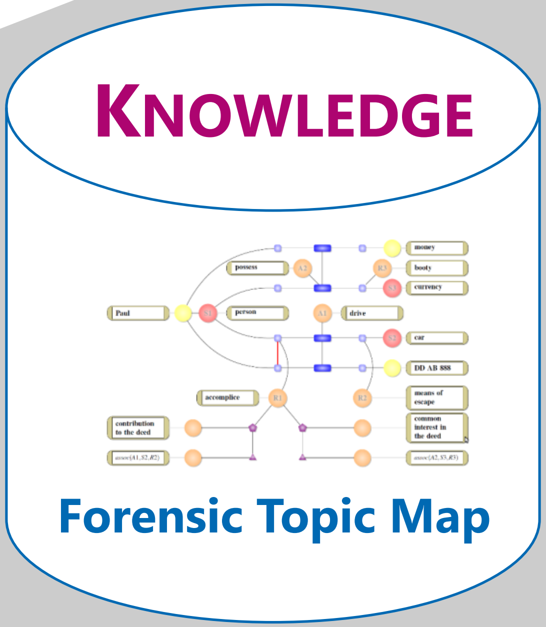
HOCHSCHULE MITTWEIDA

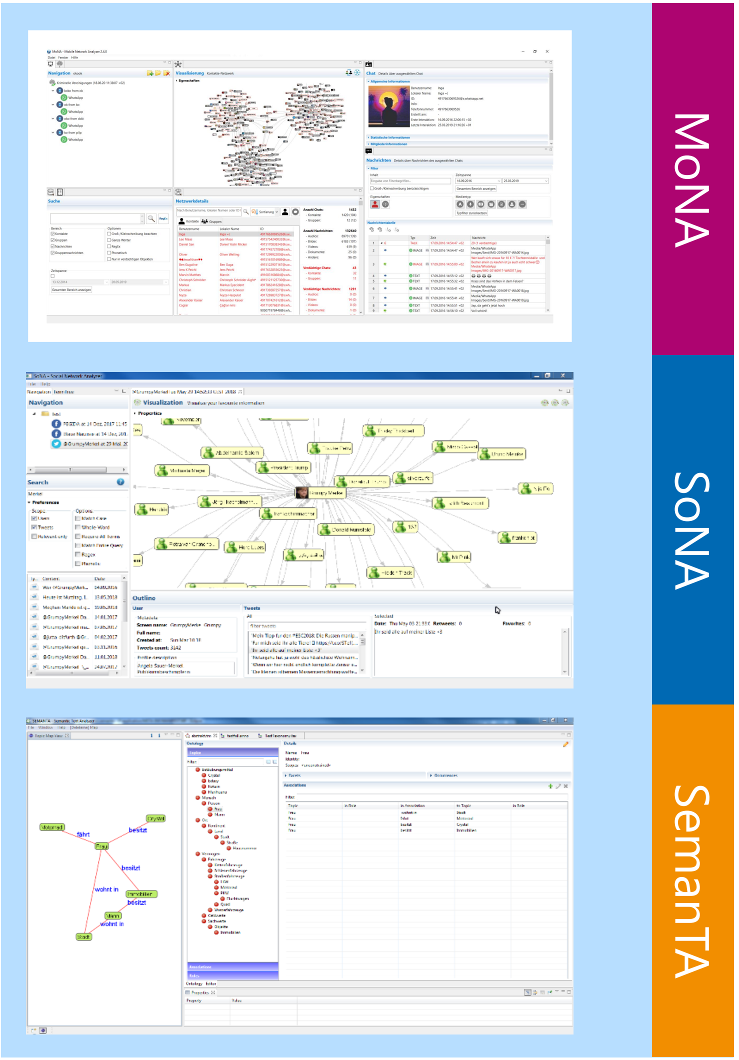
Hypotheses

Adding a knowledge model (investigative knowledge, legal norms) to text mining processes leads to comparable quality in the interdisciplinary and cross-lingual domain of forensic texts.

Forensisc Knowledge Representation as Central Element

- investigative know-ledge/experience
- legal norms



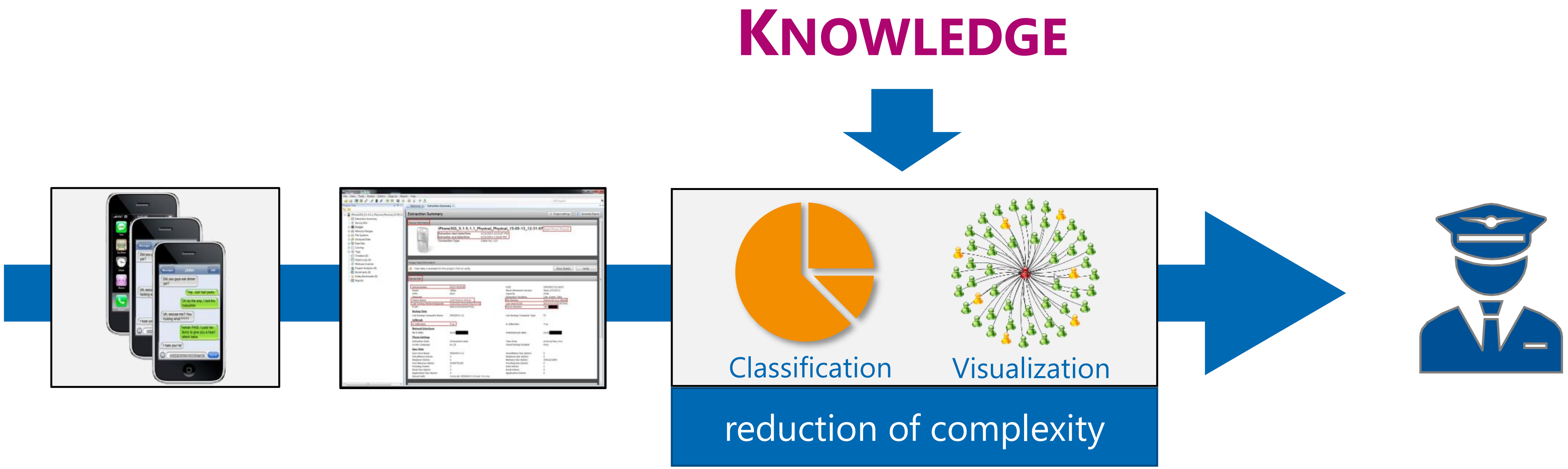


Forensics

Analysis of mobile communication

Universal Approach

- In extremely erroneous, slang and low-context texts, such as SMS, forensic information can only be detected with high reliability by **incorporating investigator knowledge**.
- An error margin can be determined.



Methods

State-of-the-Art Methods

low sensitivity

- Semi-supervised approaches
  - probabilistic language models (unigram, char-n-gram) + rules
  - performance → poor
- Difference Analysis → mainly individual spellings

low precision

- phonetic algorithms (e.g., Kölner Phonetik, Double Metaphone)

HOCHSCHULE MITTWEIDA

Hypothesis

Search space reduction and conservative word matching result in high sensitivity with acceptable precision.

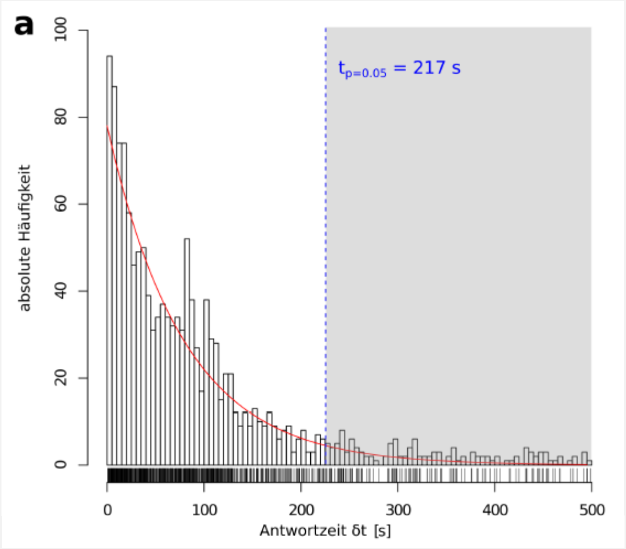
Positive side effect:

Conservation of the context →

Increase of comprehensibility

HOCHSCHULE MITTWEIDA

Method for Detection of Conversations



Frequency of response time of all messages

Assumption

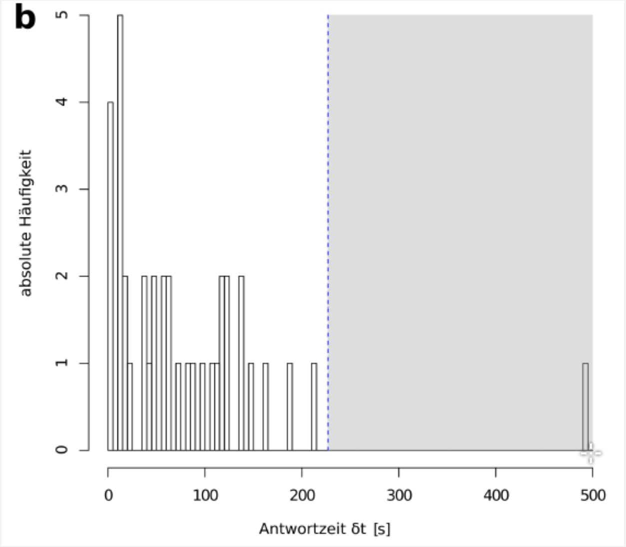
$$B_t = B_0 e^{-rt}$$

Best Fit (Regression)

$$\{r_{opt}, B_{opt}\} = \underset{r,B_0}{\operatorname{argmin}} \sum_t \left| B_t^{calc} - B_t^{opt} \right|$$

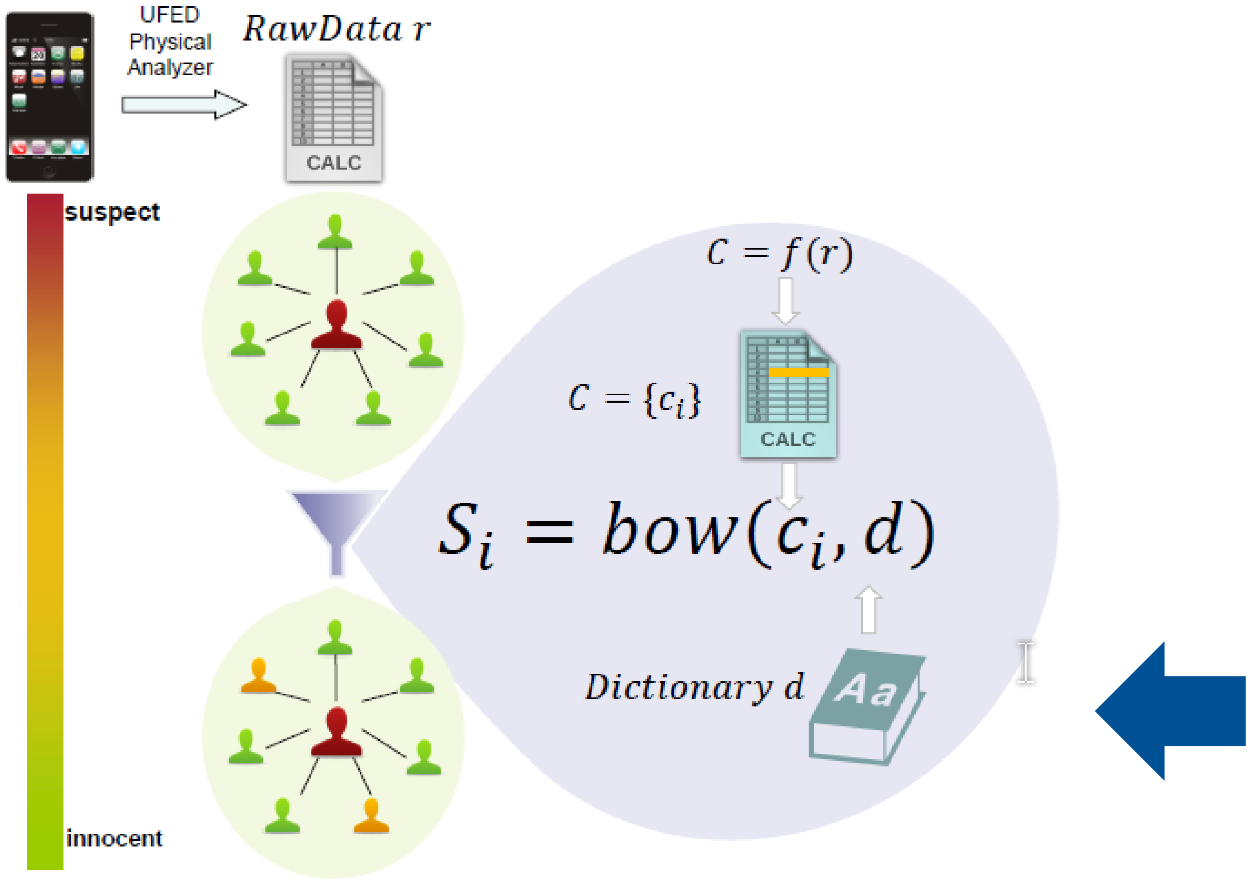
Determining Cut-Off

$$F(B) = \int_0^{t_p} B_0 e^{-rt} dt = 1 - p$$



Frequency of response times of relevant messages

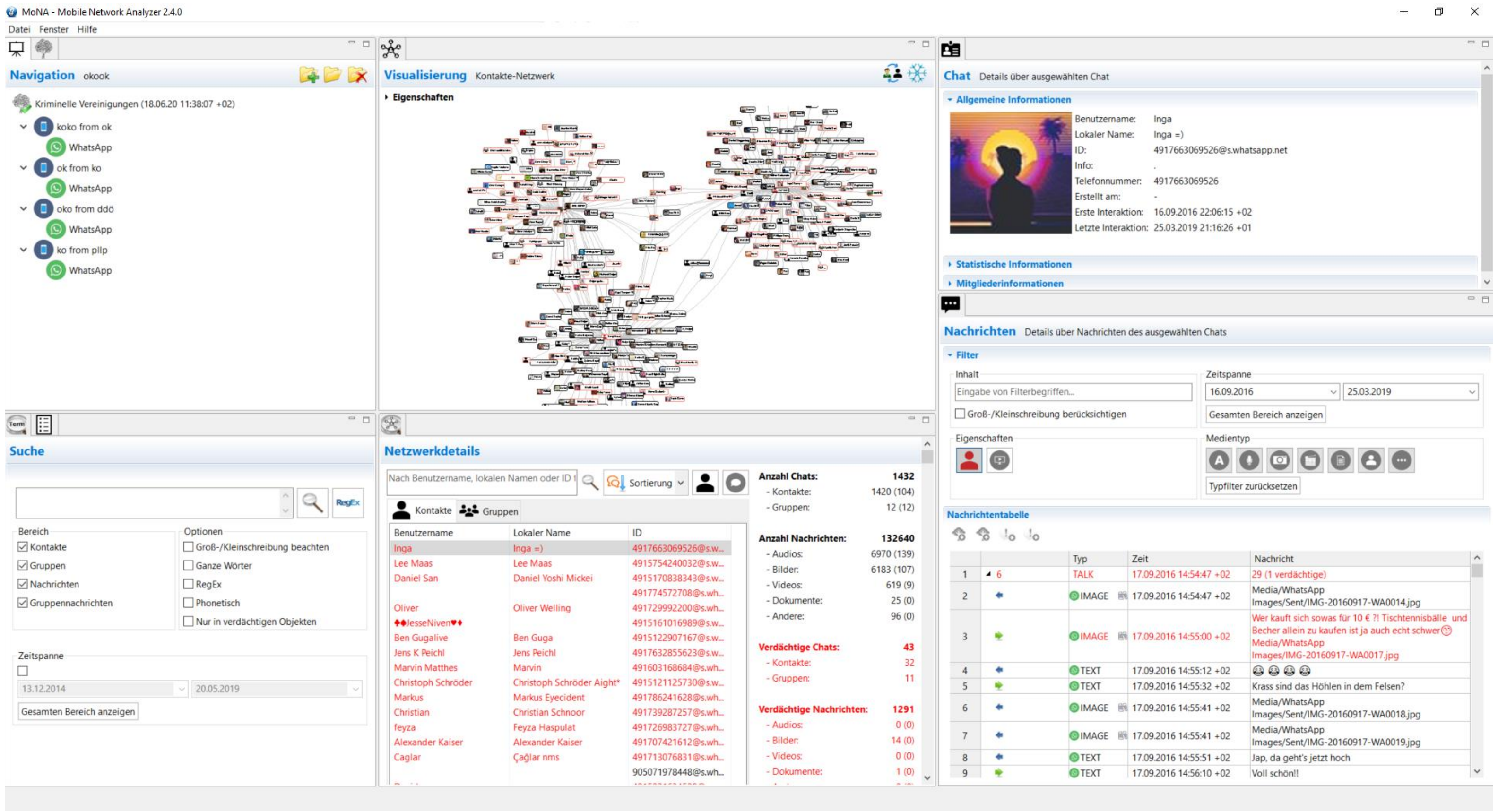
Entire Process



- semantic relations
- pattern (entities)
- Multi (cross)-linguality

HOCHSCHULE MITTWEIDA

Analysis platform for mobile communication(MoNA)



Reduction of the manual effort by >> 70%

Interactive analysis allows constant adaptation of the criminalistic hypothesis

Cross-lingual through parallel knowledge model

Prevention

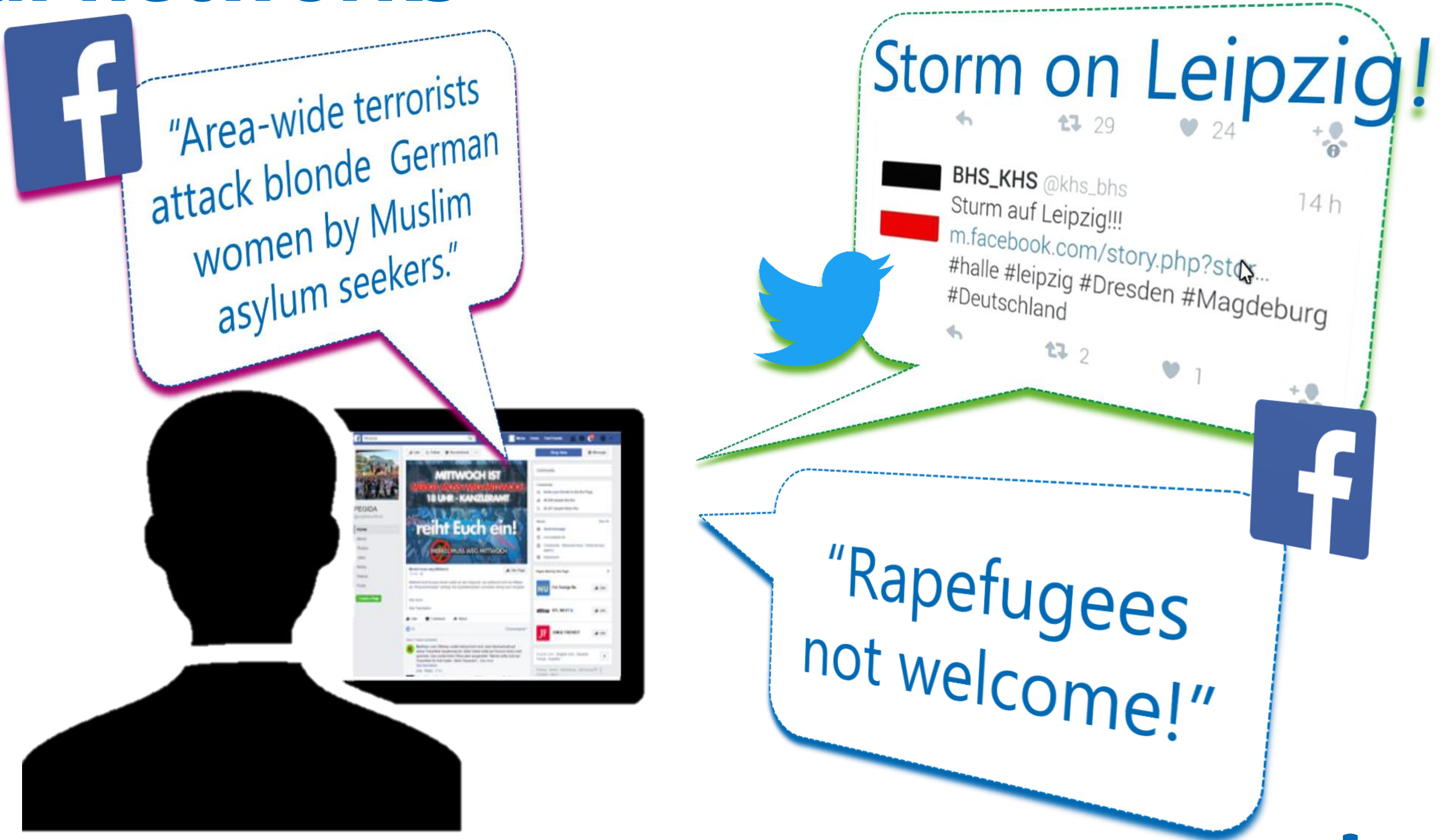
Analysis of social networks

Is crime predictable using social networks and scientific methods?

Rioting in the wake of demonstrations, sporting events or as a result of political dissatisfaction often becomes apparent in advance in the social media.

Terrorists often recruit their future assassins via social networks.Amok runners often signal their readiness in social networks.

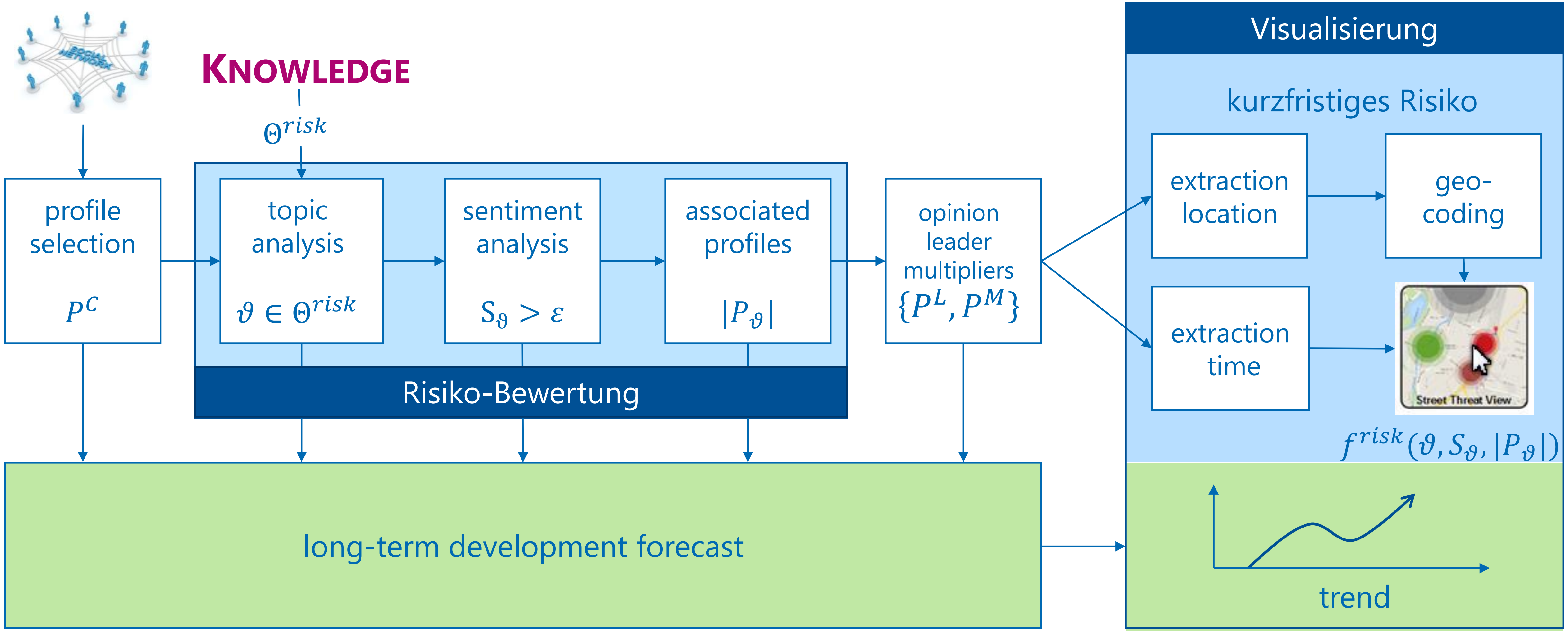
Rioters often announce themselves in social networks



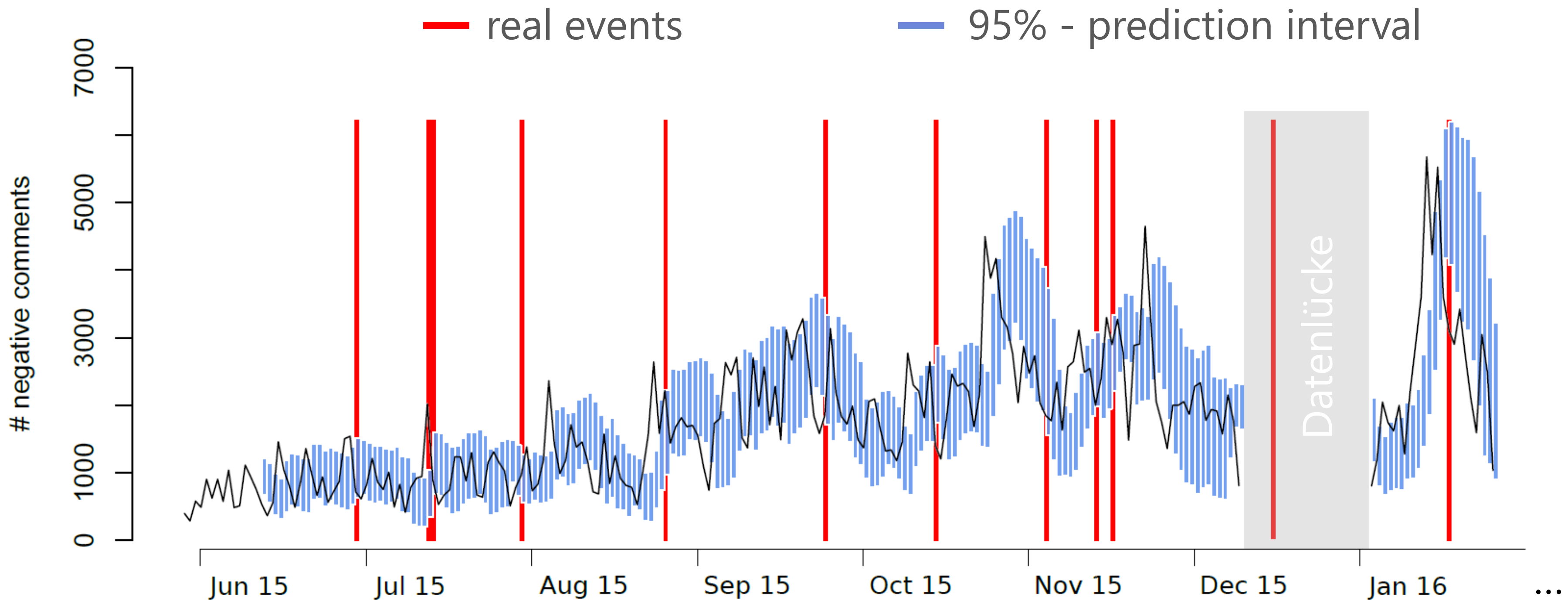
Hypothesis

By monitoring social networks, damage events in the real world can be predicted.

Process model for hazard prediction



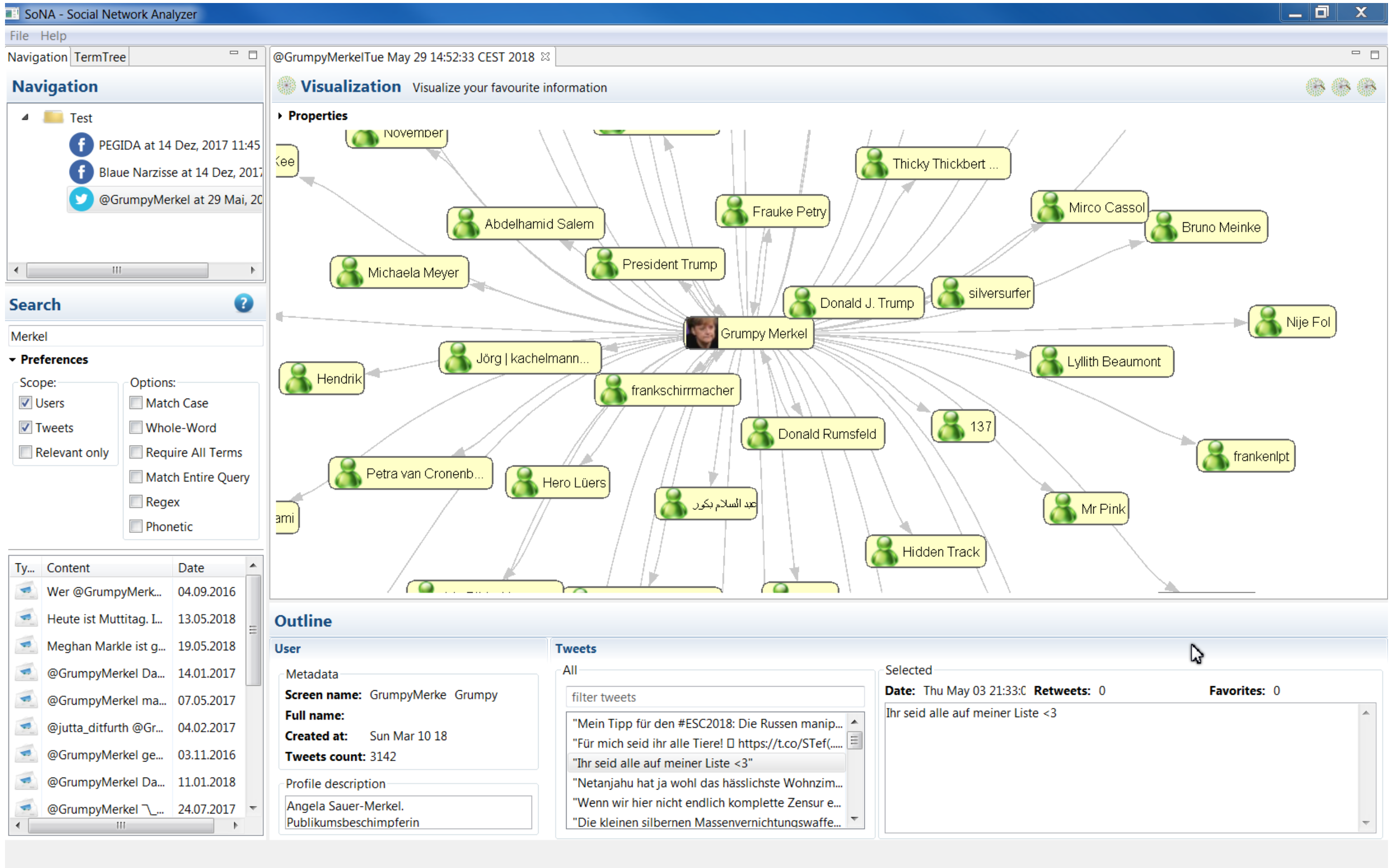
Prediction of events through sentiment analysis



Sentiment scores of the Facebook page of Pegida e.V.

Cooling phases often mark real events

Analysis Platform for Social Networks (SoNA)



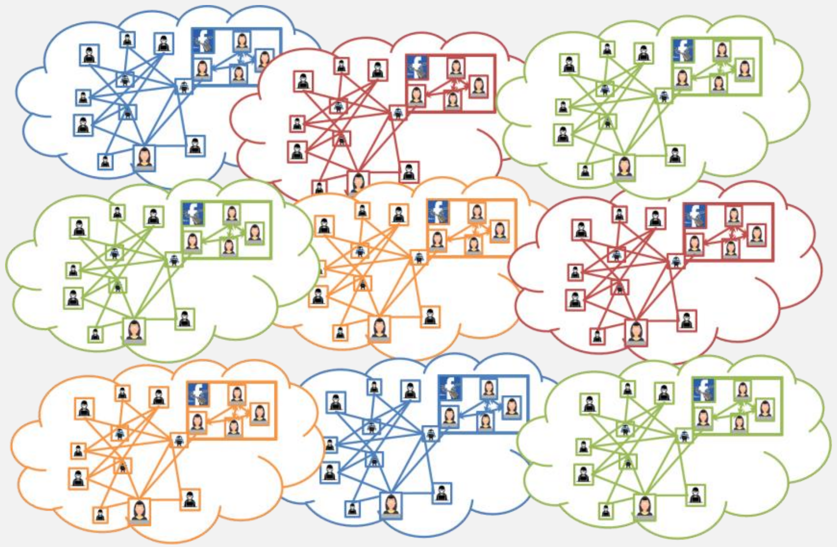
Opinion leader detection

Evaluation of the risk potential

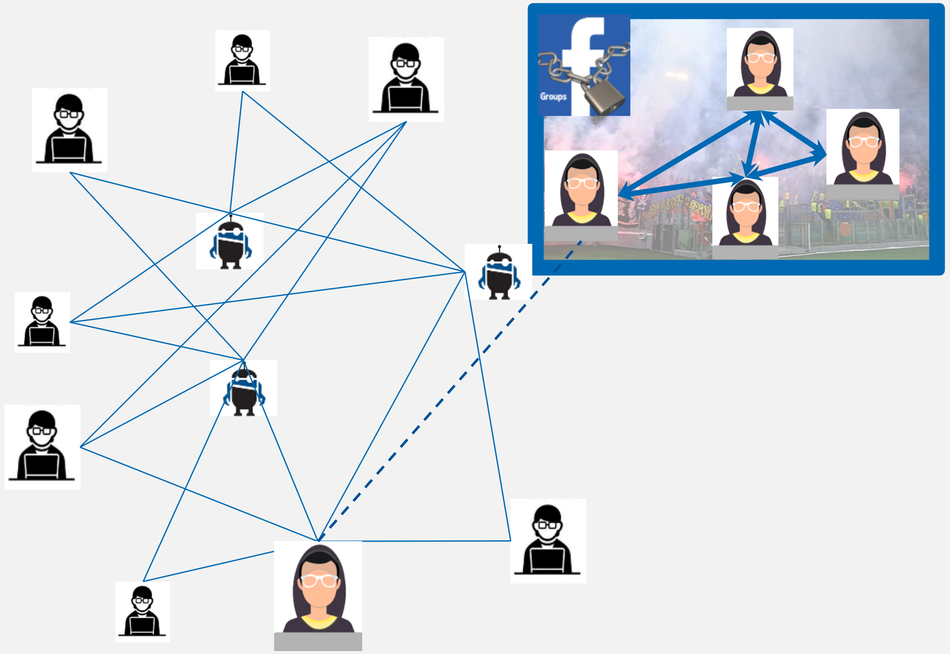
Cross-lingual through parallel knowledge model

Challenges

Huge amount of potential (hazardous) profiles



Closed/Secret groups and bots



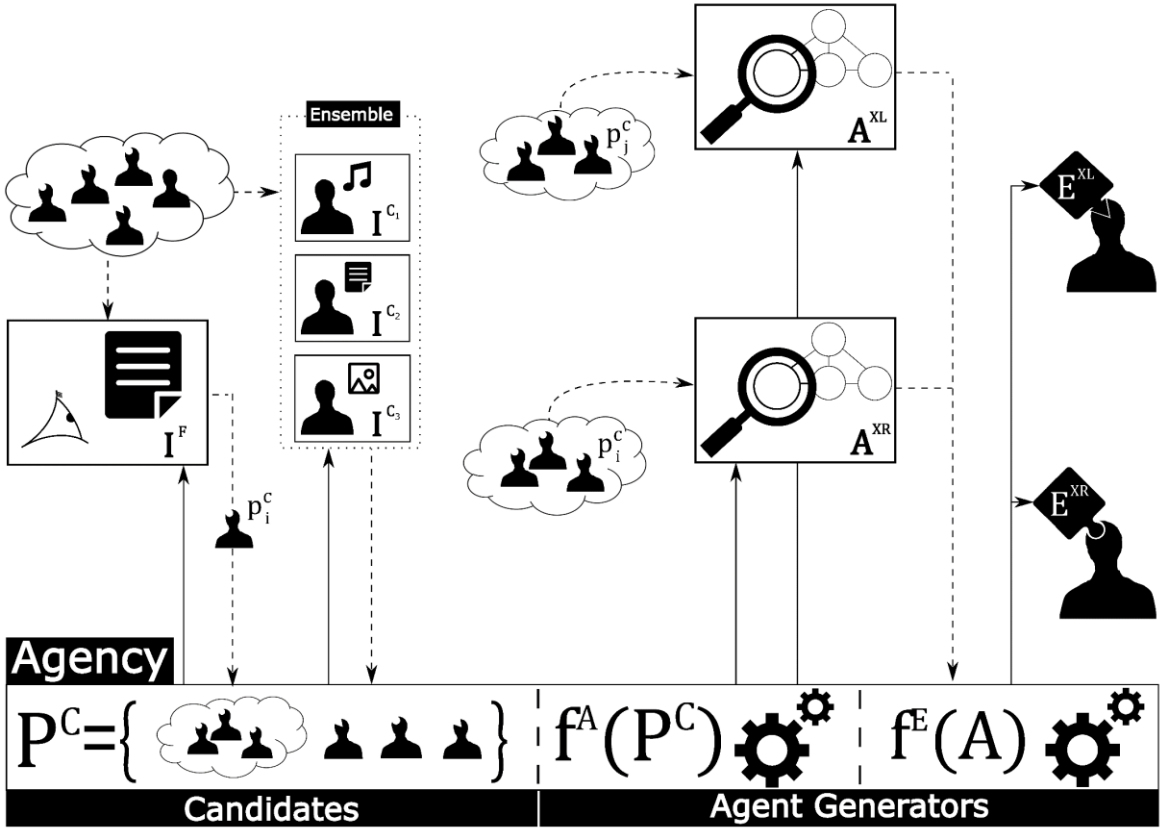
Hypothesis

By transferring strategies of the human immune system, threats in social networks can be effectively identified.

Strategies:

- pattern recognition
- adaptation

Agent-based analysis of social networks



Actors of an artificial immune system for social networks

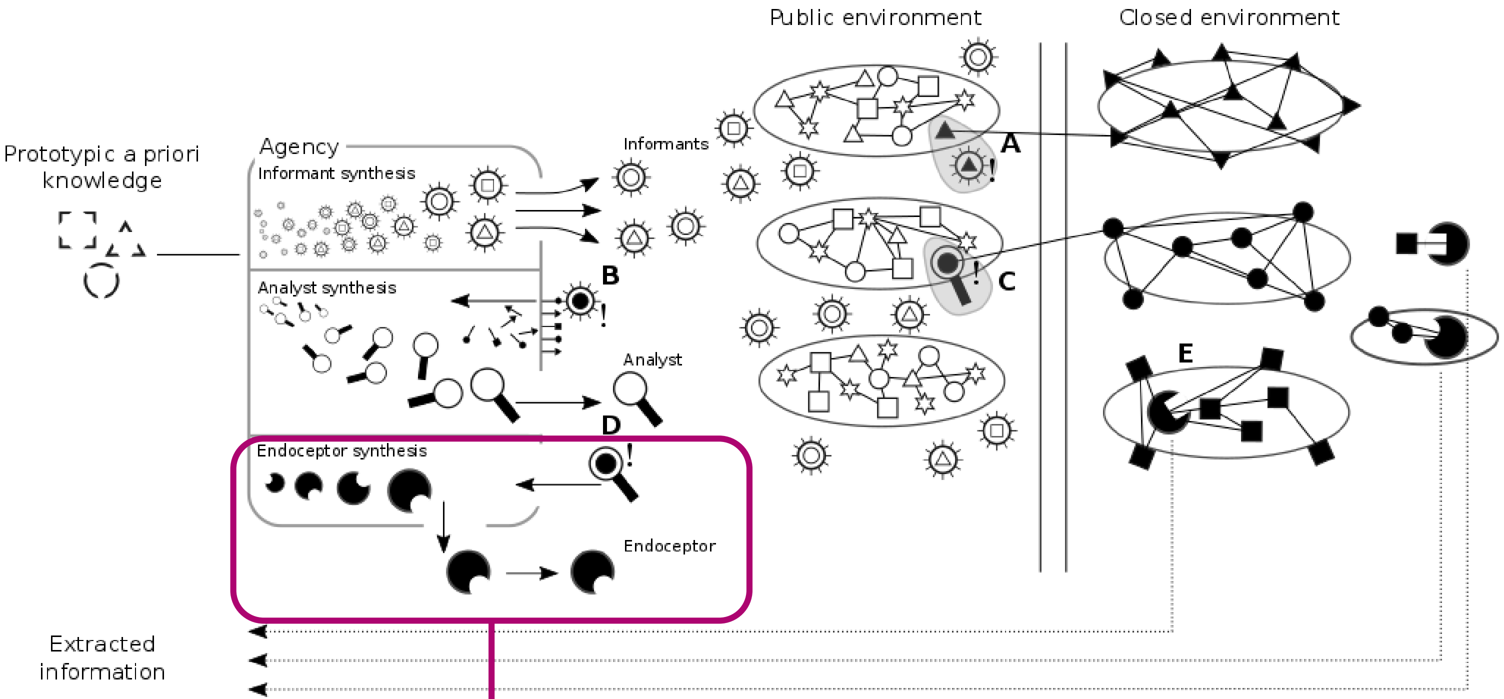
Scoring function

$$r(p_i^c) = \lambda \frac{count(I^o, p_i^c)}{\sum_{p_j \in P^c} count(I^o, p_j^c)} + (1-\lambda) \frac{1}{|I^{c_j}|} \sum_{j=1}^{|I^{c_j}|} w_j I^{c_j}(p_i^c)$$

Activation function

$$\alpha_A(p_i^c) = \begin{cases} 1, & if\ r(p_i^c) > \epsilon \\ 0, & sonst \end{cases}$$

An Artificial Immune System



Activities in an artificial immune system for social networks (process view)

Which profiles should be contacted ?

Which profile provides the **most valuable information**?

Opinion Leader

What exactly does that mean?

“Opinion leadership is the degree to which an individual is able to **influence informally** other individuals’ attitudes or **overt behavior** in a desired way with relative frequency.” [Rogers, 1962, p. 331]

What makes an influencer?

Katz and Lazarsfeld 1957 defined the following features:

(1) personification of certain values,
(2) competence,
(3) strategic position in the social network (topology).

What does "influence" mean ?

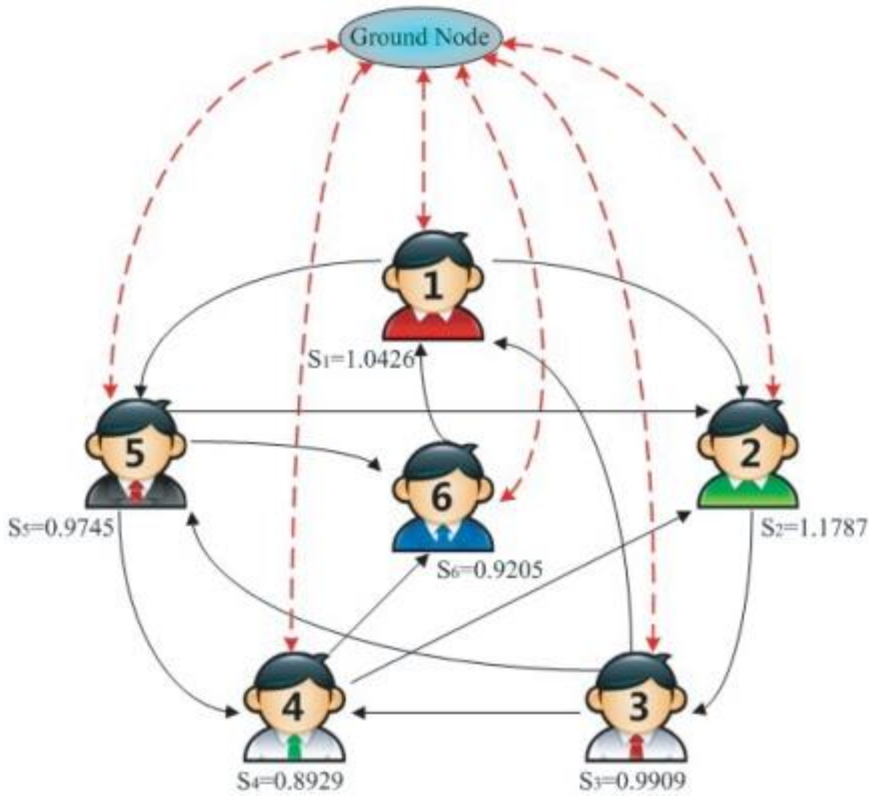
	Spreading information quickly	Write something of importance
Meaning	• depending on a strategic position in the network • own activity is the most important factor	• strategic position is mainly determined by mentions (quotations) • dependence on the topic
	Are Social Bots Influencers?	Change over time!
Approaches	• topology-based	• topology-based • content-based
Methods	• network centrality measures, PageRank, **LeaderRank**	• network centrality measures, PageRank, **LeaderRank,** sentiment analysis, topic mining

How does LeaderRank work?

- Users are nodes, directed edges connect followers with leaders
- Random walk on this graph, starting with
  $s_i(0) = 1, s_g(0) = 0$

$$s_i(t+1) = \sum_{j=1}^{N+1} \frac{a_{ij}}{k_j^{out}} s_j(t)$$

- Finds nodes that spread information further and faster.



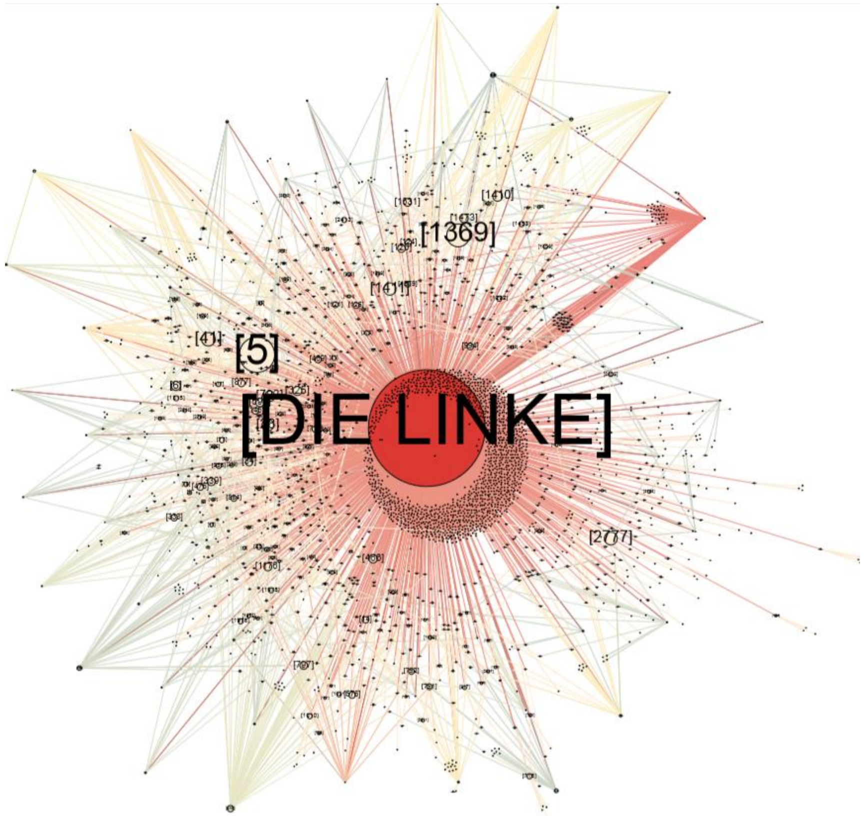
HOCHSCHULE MITTWEIDA

Problems with LeaderRank

In networks with star topology :

- the network **owner** is **highly centralized**
- high centrality of a fraction of the nodes leads to a **strongly distorted LeaderRank** distribution
- **competence** is **not considered**
- peripheral nodes are not adequately represented

➢ **LeaderRank is not meaningful!**



Facebook network: "Die Linke" over 5 months

Hypothesis

Using the normalized LeaderRank skewness, star topologies can be detected in network graphs.

HOCHSCHULE MITTWEIDA

Detection of a Star-Shaped Topology

How can the degree of approximation to the star topology be quantified?

Normalized LeaderRank-Scewness $\hat{\nu}$:

$$\nu_{LR} = \left| \frac{1}{N} \sum_{i} z(LR_i)^3 \right| \qquad \hat{\nu} = \frac{\nu - \nu_{min}}{\nu_{max} - \nu_{min}}$$

Normalized LeaderRank skewness [0,1] shows how strongly a network is distorted towards the star topology.

- $\hat{\nu}$ for regular graphs = 0
- $\hat{\nu}$ for star-shaped graphs = 1

Normalized Graph-Entropy $\widehat{H}$:

$$H = -\sum_{i=1}^{N} \frac{\deg(v_i)}{\sum_{j=1}^{N} \deg(v_j)} \log_2 \frac{\deg(v_i)}{\sum_{j=1}^{N} \deg(v_j)}$$

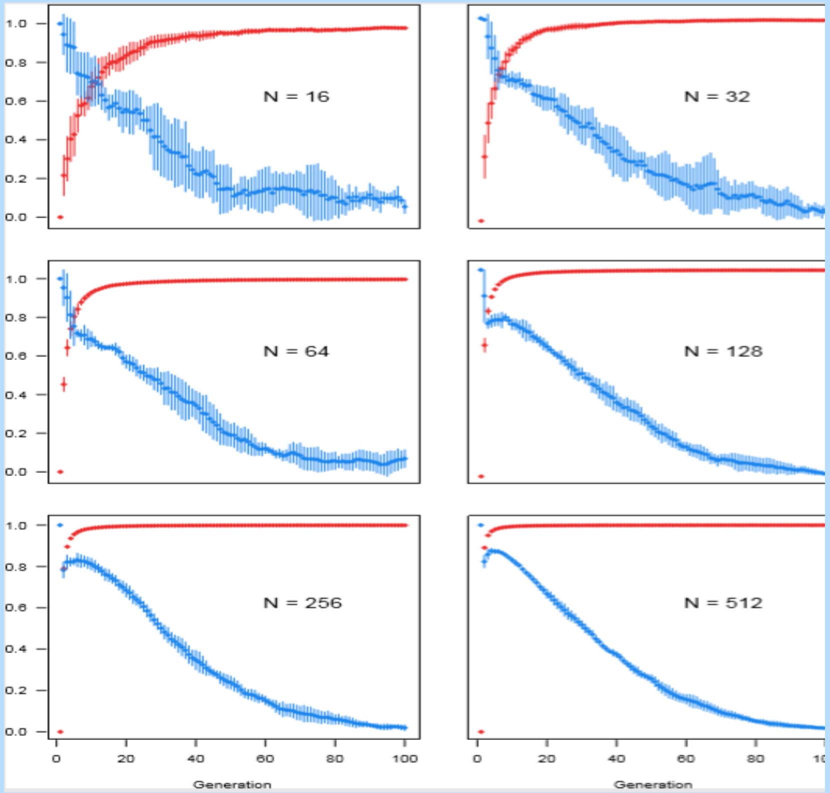
Normalized graph entropy quantifies the uncertainty of a specific path of information distribution.

- $\widehat{H}$ for regular graphs = 1
- $\widehat{H}$ for star-shaped graphs = 0

Comparison of both measures

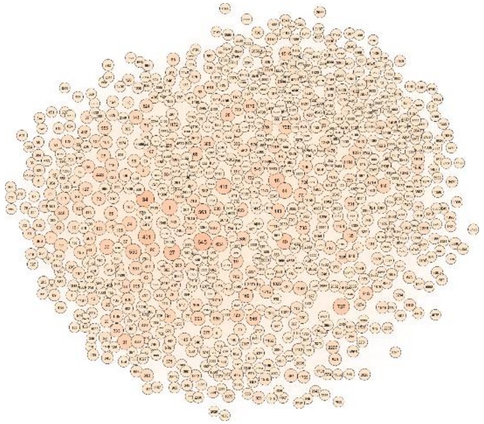
Experiment

- 6 networks with star topology with fixed number of nodes (N=16, 32, 64, 128, 256, 512)
- mutation over 100 generations towards a regular graph
- In each generation, edges are randomly added or removed between each pair of nodes



Normalized Graph-Entropy
Normalized LeaderRank-Scewness

Test on real networks

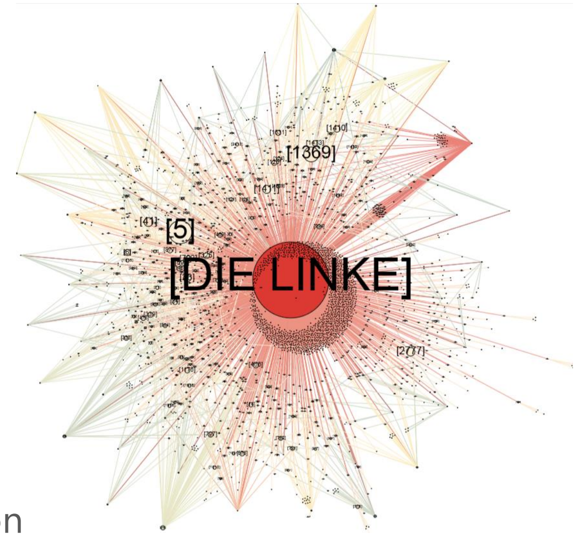


"Epinions" - Network

(almost regular)

month	actors	posts	comments	replies	$\hat{H}$	$\hat{\nu}_{LR}$
January	2,878	26	2,955	3,471	0.19	0.98
February	2,146	33	2,196	2,062	0.24	0.98
March	3,196	40	3,501	3,245	0.17	0.97
April	2,432	26	2,558	3,295	0.22	0.98
May	4,765	31	4,130	5,674	0.10	0.98
Epinions	75,879	n/a	n/a	n/a	0.65	0.07

norm. Graph-Entropy and norm. LeaderRank-Scewness in Comparison



"Die Linke" - Facebook

(almost completely star-shaped)

The normalized LeaderRank skewness, as a function of network regularity, enables a stable detection of star-shaped topologies.

Hypothesis

The irregularity of a star graph can be compensated by punishing high activity with low mentioning.

HOCHSCHULE MITTWEIDA

Way out : CompetenceRank

Variant of the LeaderRank adapted to competence

$$CR(L_i) = \frac{LR(L_i)}{1 + \frac{k_i^{out}}{k_{total}^{out}}(LR_{total} = N)}$$

In regular graphs :

$$k_i^{out} = k_j^{out} = D \forall\ (v_i, v_j)$$

$$CR(L_i) = \frac{LR(L_i)}{1 + \frac{D}{ND} N} = \frac{1}{2} LR(L_i)$$

Assumption:

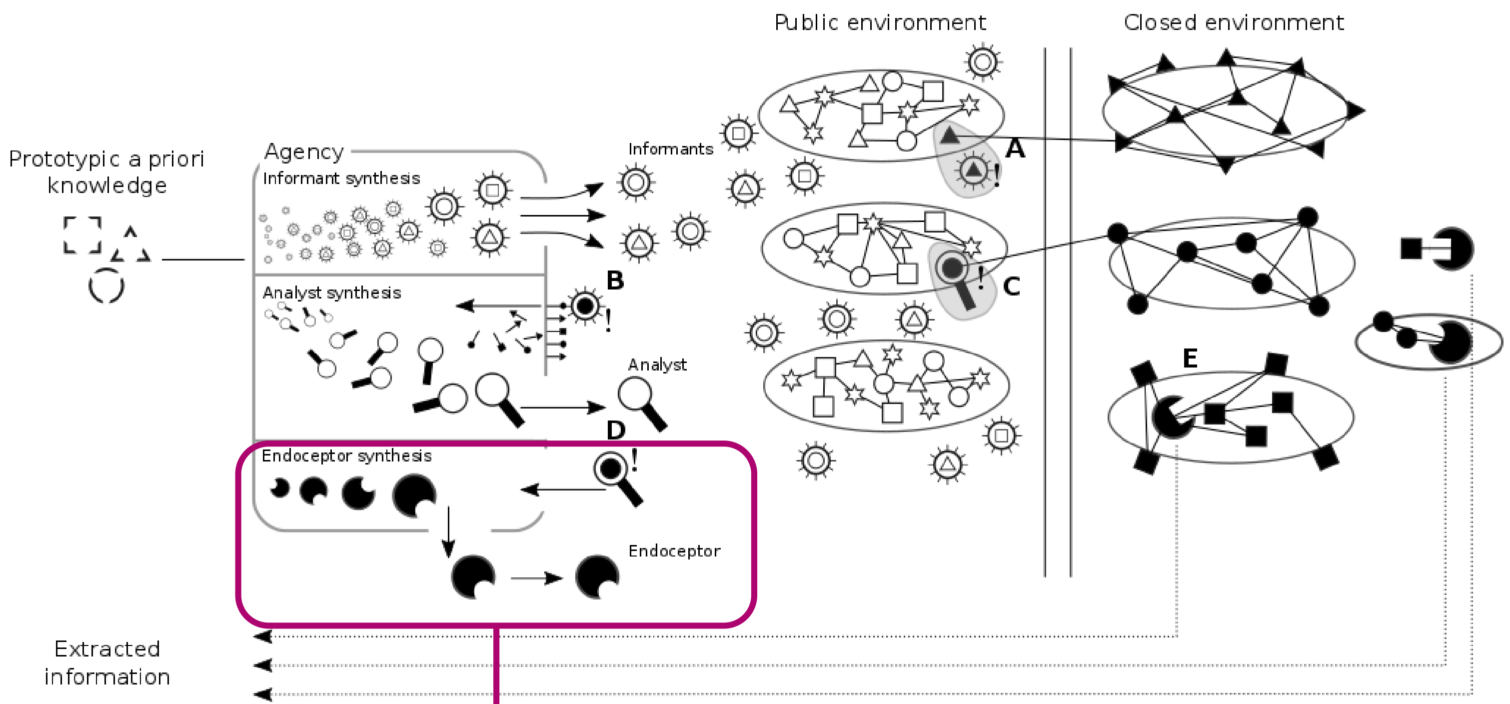
$$LR(L_i) = CR(L_i)$$

$$CR(L_i) = 2\frac{LR(L_i)}{1 + \frac{k_i^{out}}{k_{total}^{out}} N}$$

Cumulative discrepancy is a function of network regularity

$$\sum_{i=1}^{N} |CR(L_i) - LR(L_i)|$$

An Artificial Immune System



Activities in an artificial immune system for social networks (process view)

Which profiles should be contacted ?

Profiles with a high CompetenceRank!

Conclusion

- ✓ Investigator knowledge helps to improve text mining in forensics
- ✓ **MoNA** is an analysis platform for mobile communication that incorporates this paradigm.
- ✓ Algorithm for **conversation detection**
- ✓ Rating algorithm with search space reduction and conservative word matching
- ✓ With **SoNA**, an analysis platform for social networks was created incorporating this paradigm.
- ✓ Process for predicting potential hazardous events
- ✓ **Model of an artificial immune system** for social networks
  - ✓ **CompetenceRank** as an improved measure of opinion leadership

Future Work

- **Joint Semantic Analysis**: joint analysis of media and text for mobile devices
- Incorporation of context data (CPLSA/NetPLSA)
- Time related analysis of messages → **Prediction of cyclic recurring topics**
- **Evolution** of topics
- **Multilingual text analysis** with minimal amount of training data
  - approach through adaptation and expansion of **Human Behaviour-based Optimization**

Questions?

Feel free to contact me:
spranger@hs-mittweida.de

HOCHSCHULE MITTWEIDA
University of Applied Sciences