Identifying an individual by leveraging social relationships

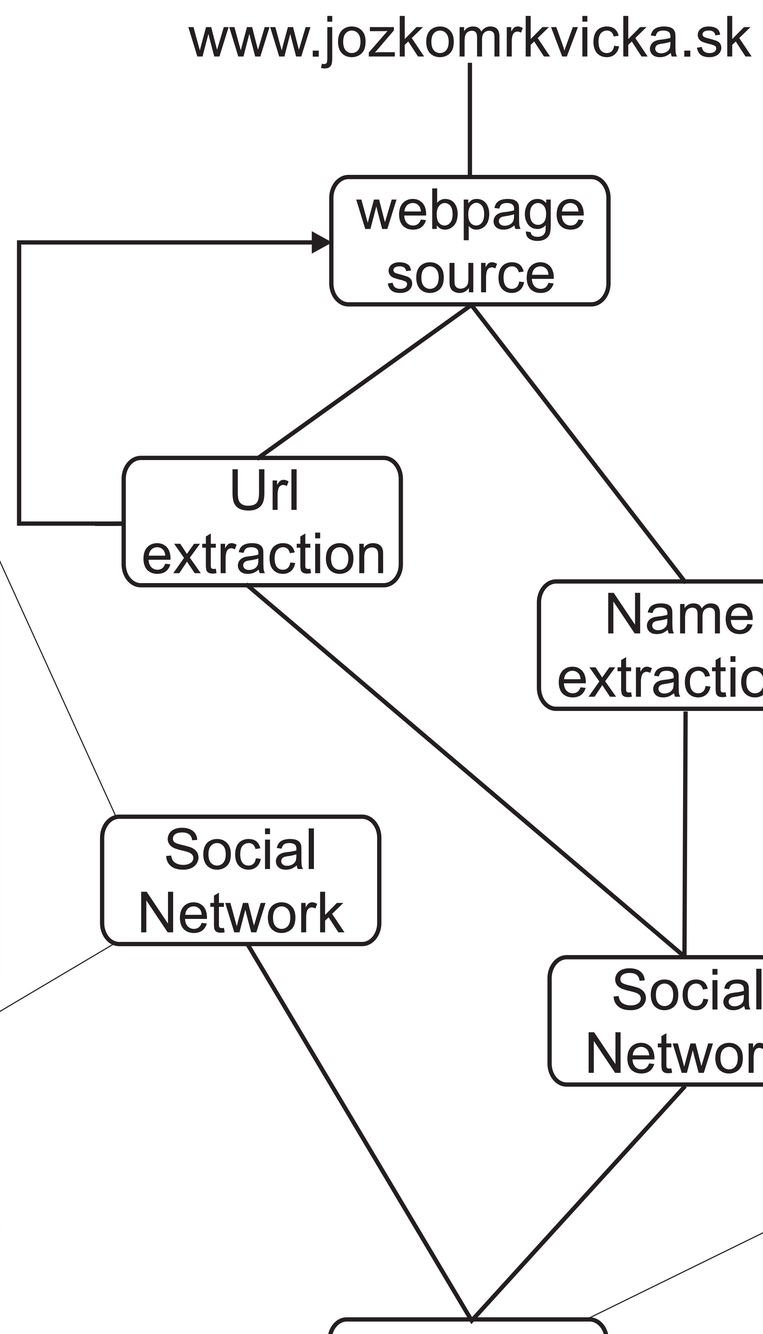
Goals

- identify **information** related to particular person
- identify **duplications** in social network
- distinguish **namesakes**
- identify **name varians** of person

Background knowledge

- some information about person
- help identify individual
- distinguish namesakes

- our BK is person's social network



Process of identification

1. Inputs - webpage url and **background knowledge**
2. Source processing - **Url** and **Name extraction**
3. Social Network creation
4. Social **Network comparison**

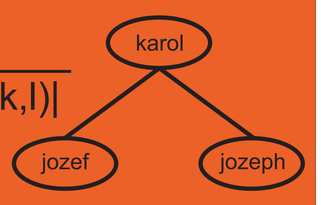
Name extraction

- dictionary of english first names
- regular expression
- 4 formats
  - Hanna Maden
  - Hanna S. Maden
  - Hanna Sofia Maden
  - Maden Hanna S.

Connected triple

basic

$$sim(i,j) = \frac{|CT(i,j)|}{\max_{k=0..m,l=0..n} |CT(k,l)|}$$



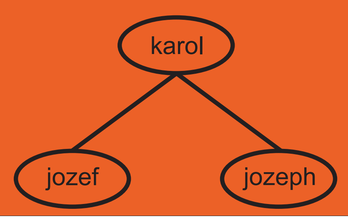
modified

$$sim(i,j) = \frac{|CT(i,j)|}{\max_{k=i \vee j,l=0..n} |CT(k,l)|}$$

Social Network Comparison

- Levenshtein method

- Modified Connected triple



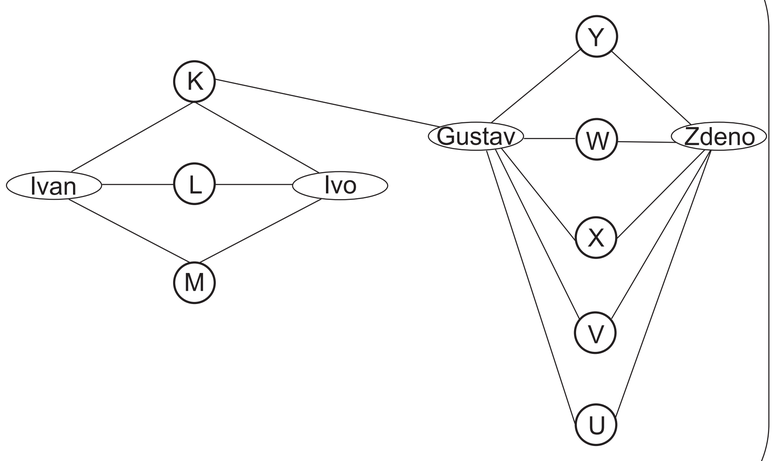
Name extraction

1. experiment - extract 1. format

	Mean	Standard deviation	Min	Max
Precision	88,6%	9,1	73,7%	95,2%
Recall	46,1%	11,1	29,6%	56,4%

2. experiment - extract all 4 formats

	Mean	Standard deviation	Min	Max
Precision	95,8%	2,6	94,7%	98,4%
Recall	51,3%	12,2	31,6%	59,8%



Network comparison

	Havrans	Novaks
# vertices	329	293
# edges	610	526
lev. 85% + same address	28	25
lev. 85% + CT	28	23
lev. 85% + CT + same address	23	15
precision	82,2%	65,2%
recall	82,2%	60%

References:

Bekkerman, R., McCallum, A.: Disambiguating Web appearances of People in a Social Network. In: WWW '05, ISBN: 1-59593-046-9, ACM, New York, NY, USA, 2005, pp. 463-470.
Mann G. S., Yarowsky D.: Unsupervised Personal Name Disambiguation, Proceedings of the seventh conference on Natural language learning at HLT-NAACL, 2003, pp. 33 - 40
Reuther P., Walter B., Ley M., Weber A. and Klink S - Managing the Quality of Person Names in DBLP. ECDL 2006. LNCS 4172. pp. 508-511

IITSRC 2009
Bc. Katarína Kostková
Ing. Michal Barla