Towards a Free Social Network Webcrawler
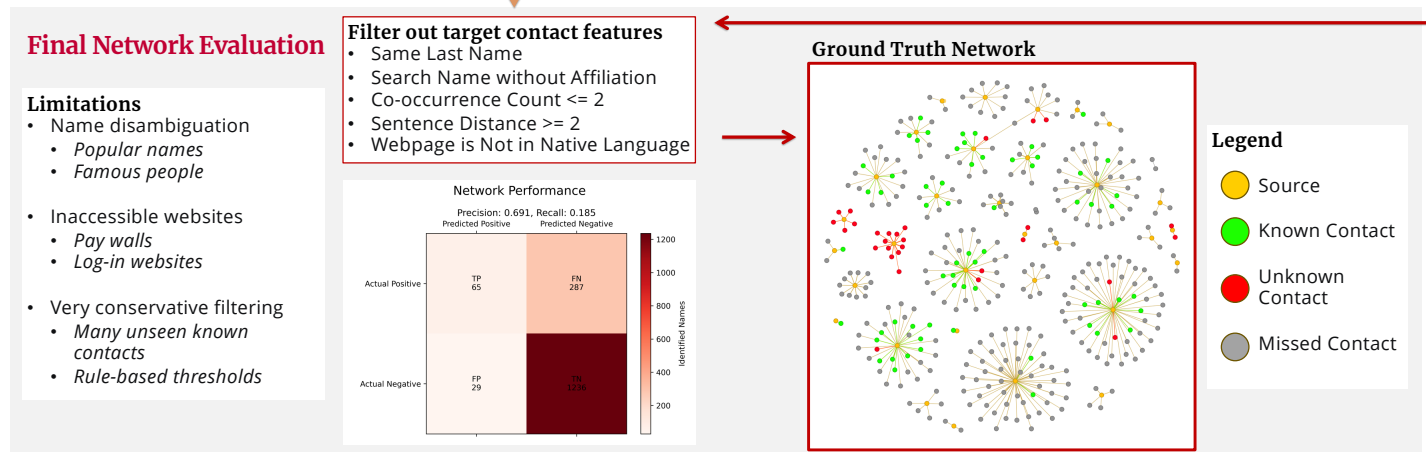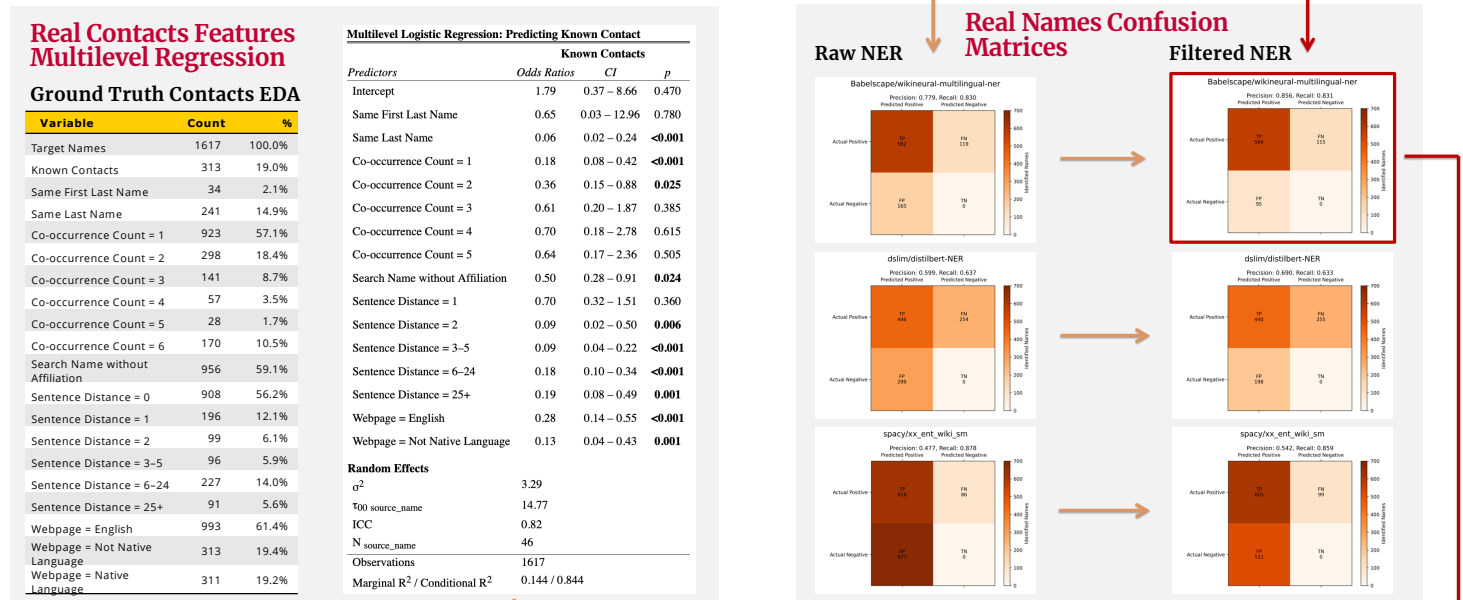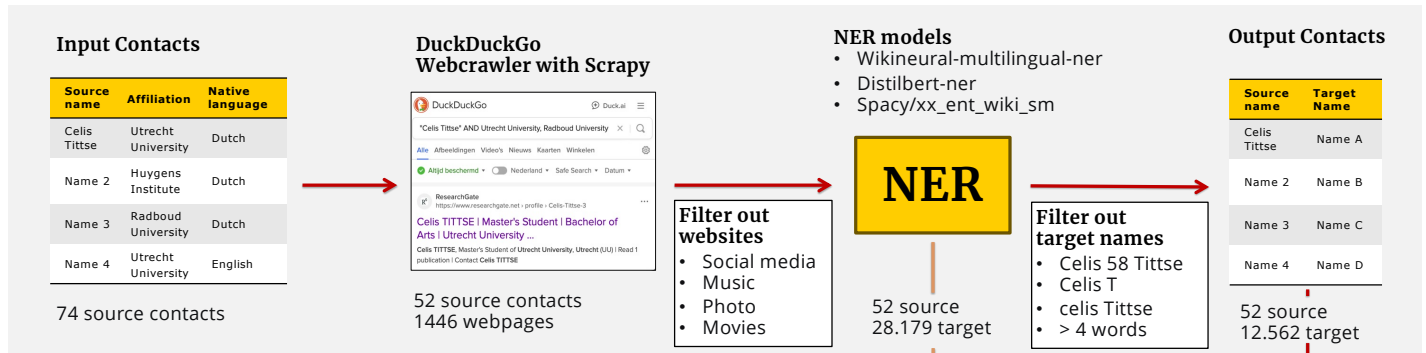## The verification of real connections from webpages

Celis Tittse

Social Network Crawler Pipeline

Input Contacts

Source name	Affiliation	Native language
Celis Tittse	Utrecht University	Dutch
Name 2	Huygens Institute	Dutch
Name 3	Radboud University	Dutch
Name 4	Utrecht University	English

74 source contacts

DuckDuckGo Webcrawler with Scrapy

DuckDuckGo
"Celis Tittse" AND Utrecht University, Radboud University

ResearchGate
https://www.researchgate.net › profile › Celis-Tittse-3
Celis TITTSE | Master's Student | Bachelor of Arts | Utrecht University ...
Celis TITTSE, Master's Student of Utrecht University, Utrecht (UU) | Read 1 publication | Contact Celis TITTSE

52 source contacts
1446 webpages

Filter out websites
- Social media
- Music
- Photo
- Movies

NER models
- Wikineural-multilingual-ner
- Distilbert-ner
- Spacy/xx_ent_wiki_sm

NER

52 source
28.179 target

Filter out target names
- Celis 58 Tittse
- Celis T
- celis Tittse
- > 4 words

Output Contacts

Source name	Target Name
Celis Tittse	Name A
Name 2	Name B
Name 3	Name C
Name 4	Name D

52 source
12.562 target

Real Contacts Features Multilevel Regression

Ground Truth Contacts EDA

Variable	Count	%
Target Names	1617	100.0%
Known Contacts	313	19.0%
Same First Last Name	34	2.1%
Same Last Name	241	14.9%
Co-occurrence Count = 1	923	57.1%
Co-occurrence Count = 2	298	18.4%
Co-occurrence Count = 3	141	8.7%
Co-occurrence Count = 4	57	3.5%
Co-occurrence Count = 5	28	1.7%
Co-occurrence Count = 6	170	10.5%
Search Name without Affiliation	956	59.1%
Sentence Distance = 0	908	56.2%
Sentence Distance = 1	196	12.1%
Sentence Distance = 2	99	6.1%
Sentence Distance = 3–5	96	5.9%
Sentence Distance = 6–24	227	14.0%
Sentence Distance = 25+	91	5.6%
Webpage = English	993	61.4%
Webpage = Not Native Language	313	19.4%
Webpage = Native Language	311	19.2%

Multilevel Logistic Regression: Predicting Known Contact

Predictors	Known Contacts		
	Odds Ratios	CI	p
Intercept	1.79	0.37 – 8.66	0.470
Same First Last Name	0.65	0.03 – 12.96	0.780
Same Last Name	0.06	0.02 – 0.24	**<0.001**
Co-occurrence Count = 1	0.18	0.08 – 0.42	**<0.001**
Co-occurrence Count = 2	0.36	0.15 – 0.88	**0.025**
Co-occurrence Count = 3	0.61	0.20 – 1.87	0.385
Co-occurrence Count = 4	0.70	0.18 – 2.78	0.615
Co-occurrence Count = 5	0.64	0.17 – 2.36	0.505
Search Name without Affiliation	0.50	0.28 – 0.91	**0.024**
Sentence Distance = 1	0.70	0.32 – 1.51	0.360
Sentence Distance = 2	0.09	0.02 – 0.50	**0.006**
Sentence Distance = 3–5	0.09	0.04 – 0.22	**<0.001**
Sentence Distance = 6–24	0.18	0.10 – 0.34	**<0.001**
Sentence Distance = 25+	0.19	0.08 – 0.49	**0.001**
Webpage = English	0.28	0.14 – 0.55	**<0.001**
Webpage = Not Native Language	0.13	0.04 – 0.43	**0.001**

Random Effects

σ^2	3.29
$\tau_{00 \ source\_name}$	14.77
ICC	0.82
N_{source_name}	46
Observations	1617
Marginal R^2 / Conditional R^2	0.144 / 0.844

Real Names Confusion Matrices

Raw NER

Babelscape/wikineural-multilingual-ner
Precision: 0.779, Recall: 0.830

dslim/distilbert-NER
Precision: 0.599, Recall: 0.637

spacy/xx_ent_wiki_sm
Precision: 0.477, Recall: 0.878

Filtered NER

Babelscape/wikineural-multilingual-ner
Precision: 0.856, Recall: 0.831

dslim/distilbert-NER
Precision: 0.690, Recall: 0.633

spacy/xx_ent_wiki_sm
Precision: 0.542, Recall: 0.859

Final Network Evaluation

Limitations
- Name disambiguation
  - *Popular names*
  - *Famous people*
- Inaccessible websites
  - *Pay walls*
  - *Log-in websites*
- Very conservative filtering
  - *Many unseen known contacts*
  - *Rule-based thresholds*

Filter out target contact features
- Same Last Name
- Search Name without Affiliation
- Co-occurrence Count <= 2
- Sentence Distance >= 2
- Webpage is Not in Native Language

Network Performance
Precision: 0.691, Recall: 0.185

Ground Truth Network

Legend
- 🟡 Source
- 🟢 Known Contact
- 🔴 Unknown Contact
- ⚪ Missed Contact

References
1. M. Nasution. (2018 – 2021). Theoretical background and problems of SMA.
2. Y. Jin, Y. Matsuo & M. Ishizuka. (2006). Applying a threshold on source to target prompt SMA.
3. G. Geleijnse & J. Korst. (2007). Scraping unknown names from biographies to construct a historical network.

uu.nl/

The information has been compiled with the utmost care but no rights can be derived from its contents.