Mining bipartite graphs to improve semantic pedophile activity detection (short paper)

R. Fournier, M. Danisch

L2TI / Institut Galilée
Université Paris-Nord, Sorbonne Paris-Cité

LIP6
CNRS et Université Pierre et Marie Curie

May 28th, 2014

Context

Paedophile activity in P2P systems

- Children victimization
- Danger for innocent users
- Policy making issues

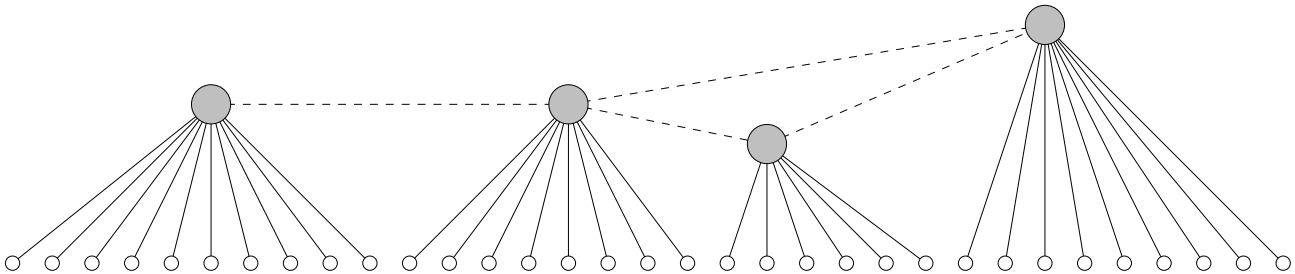
Recent research

- Identification of large file providers
- Collection of large sets of queries
- Design and validation of a detection tool [IPM 2012]

Extend this effort

Datasets

- Queries submitted to eDonkey search engine



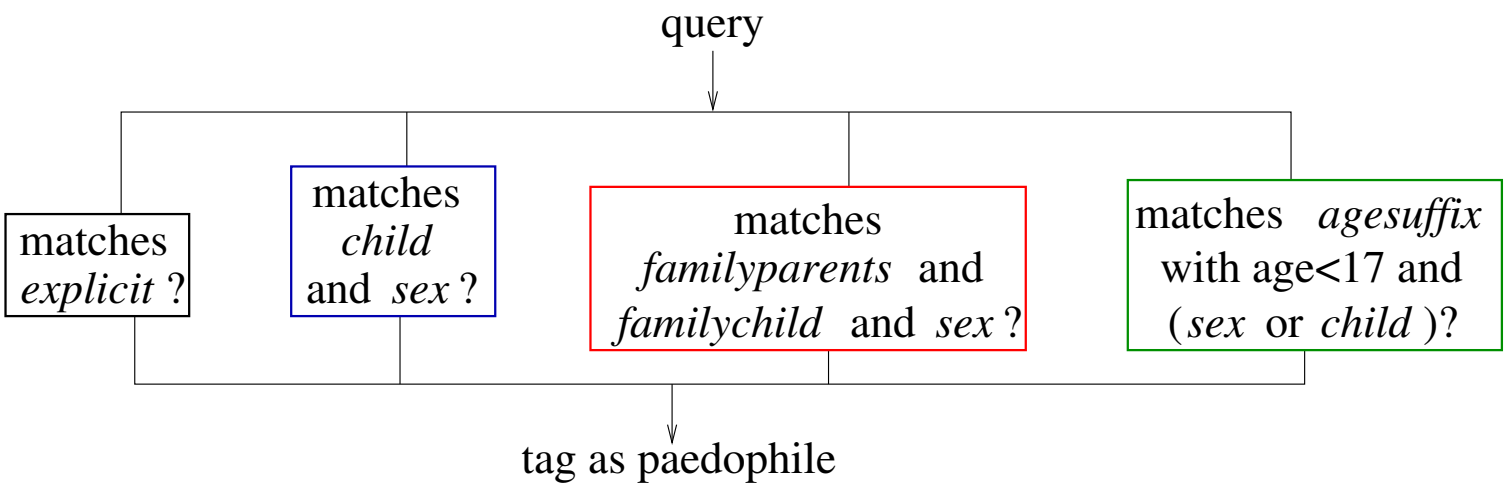
2007 10 weeks, 100 millions queries, 24 million IP addresses

- Set of queries : $q_i = (t, u, k_1, k_2, \ldots, k_n)$
  - t timestamp
  - u user information (IP address, connection port)
  - $(k_1, k_2, \ldots, k_n)$ sequence of keywords

Duly anonymised

Pedophile detection tool

- 4 semantic categories of paedophile queries



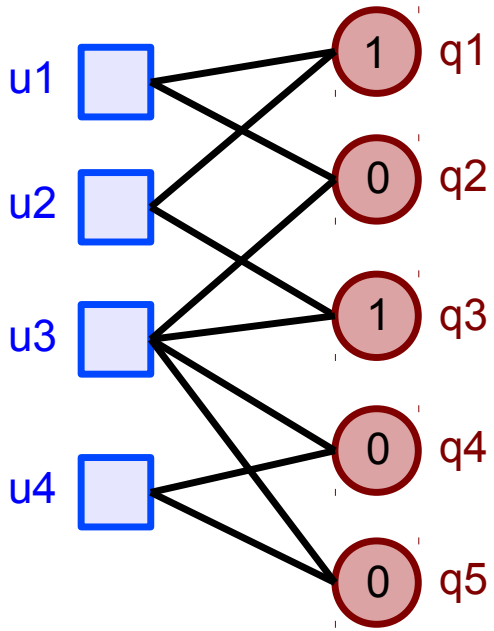
- False positives (``*sexy daddy destinys child*'')
- False negatives (``*pjk 12yo*'')

- Focus on reduced false positives rate ($< 1.4\%$)
- False negatives rate: 24.5%

Our approach

Goals

- Reduce the number of queries to process manually
- Validate existing classification



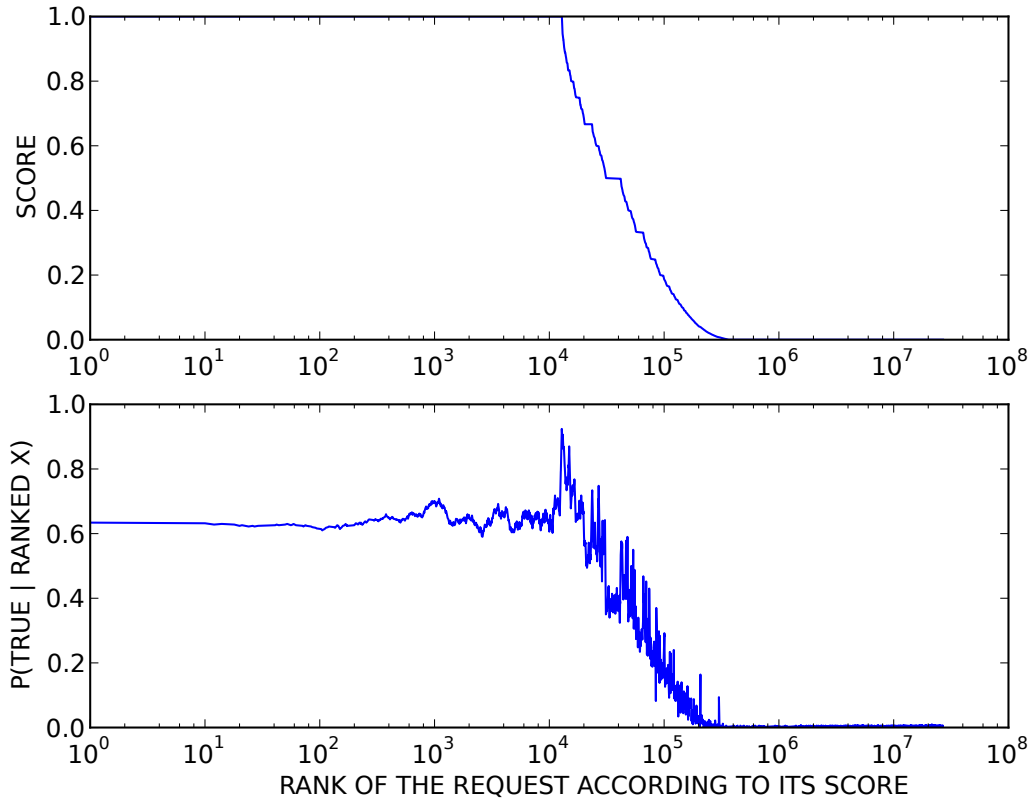
Bipartite graphs and communities

$$s_C(r) = \frac{\sum_{u \in V(r)} |C \cap R(u) \setminus \{r\}|}{\sum_{u \in V(r)} |R(u) \setminus \{r\}|}$$

$$s_1(q_2) = 0.5$$

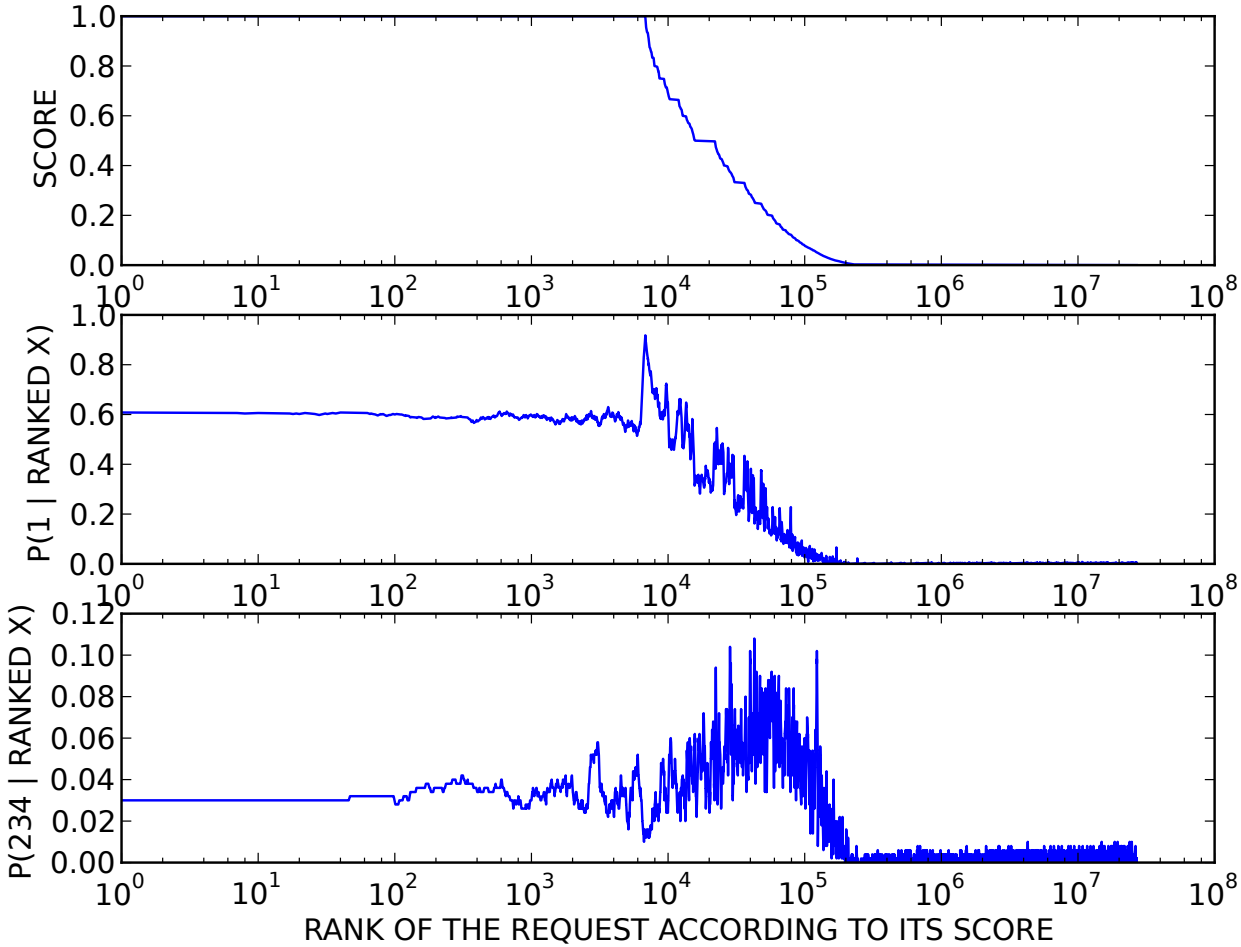
$$s'_C(r) = \frac{1}{|V(r)|} \sum_{u \in V(r)} \left| \frac{C \cap R(u) \setminus \{r\}}{R(u) \setminus \{r\}} \right|$$

Results



- 4,518 queries (out of 12,858) with score 1 not detected
- new keywords and combinations obtained
  - further analysis required to avoid increased FP rate

Results



- categories 2,3 and 4 fewly connected with category 1

Conclusion

- Measure of topological similarity between queries
- Limitation of the number of errors to process manually
- Semantic and topological categories seem linked

Future work

- Explore other scoring functions
- Explore local community completion methods
- Update the original filter by refining its lists of keywords
  - introduce new categories
  - subdivide existing categories

Thank you for your attention.

Questions?

raphael.fournier@lip6.fr