MAPAP

Measurement and Analysis of P2P Activity Against Paedophile Content

Raphaëlle Nollez-Goldbach
Project Administrator - CNRS

Safer Internet Plus

`http://antipaedo.lip6.fr`

October 2007 – October 2009

Project Presentation

Project Presentation

Partners:

- CNRS (France), Computer Science
- University College Cork (Ireland), Psychology
- University of Ljubljana (Slovenia), Social Sciences
- Nobody's Children Foundation (Poland), NGO

Project Presentation

Partners:

- CNRS (France), Computer Science
- University College Cork (Ireland), Psychology
- University of Ljubljana (Slovenia), Social Sciences
- Nobody's Children Foundation (Poland), NGO

Main goals:

- Methods and tools to protect peer-to-peer users
- Accurate information on paedophile exchanges
- Helping law enforcement institutions and NGOs

Project Presentation

Partners:

- CNRS (France), Computer Science
- University College Cork (Ireland), Psychology
- University of Ljubljana (Slovenia), Social Sciences
- Nobody's Children Foundation (Poland), NGO

Main goals:

- Methods and tools to protect peer-to-peer users
- Accurate information on paedophile exchanges
- Helping law enforcement institutions and NGOs

Actions:

- Large-scale measurements
- Analysis of P2P exchanges

Outline

1. P2P System: eDonkey
2. Measurement
3. Data Processing

1. P2P System: eDonkey

$$
\begin{array}{ccccc}
Peer & \longrightarrow & Keywords & \longrightarrow & Server \\
Server & \longrightarrow & File & \longrightarrow & Peer \\
Peer & \longrightarrow & File_{id} & \longrightarrow & Server \\
Server & \longrightarrow & List\ of\ peers & \longrightarrow & Peer
\end{array}
$$

1. P2P System: eDonkey

$$\textit{Peer} \longrightarrow \textit{Keywords} \longrightarrow \textit{Server}$$

1. P2P System: eDonkey

$$\begin{array}{ccccc} \textit{Peer} & \longrightarrow & \textit{Keywords} & \longrightarrow & \textit{Server} \\ \textit{Server} & \longrightarrow & \textit{File} & \longrightarrow & \textit{Peer} \end{array}$$

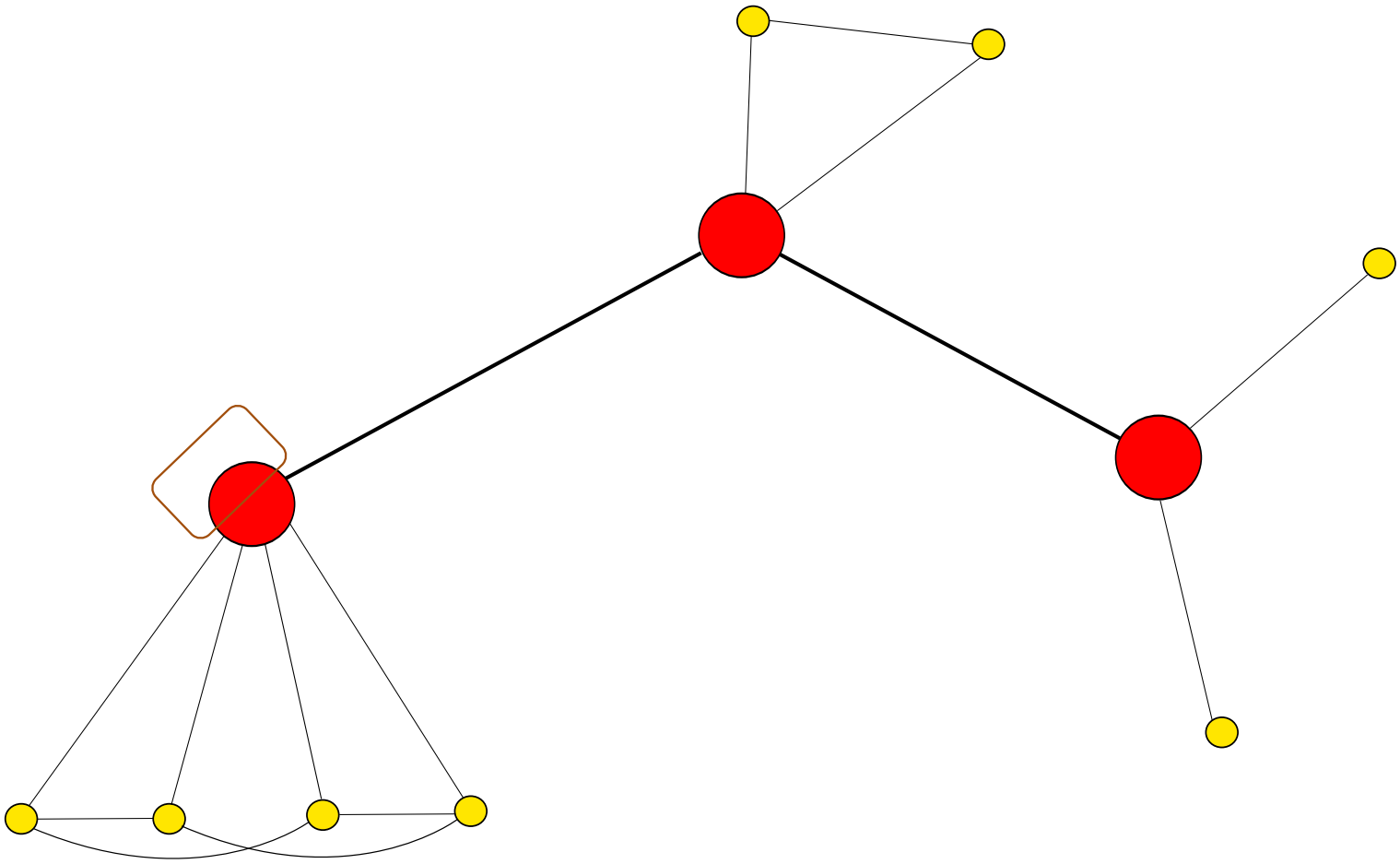
1. P2P System: eDonkey

$$
\begin{array}{ccccc}
\mathit{Peer} & \longrightarrow & \mathit{Keywords} & \longrightarrow & \mathit{Server} \\
\mathit{Server} & \longrightarrow & \mathit{File} & \longrightarrow & \mathit{Peer} \\
\mathit{Peer} & \longrightarrow & \mathit{File}_{id} & \longrightarrow & \mathit{Server}
\end{array}
$$

1. P2P System: eDonkey

$$
\begin{array}{ccccc}
Peer & \longrightarrow & Keywords & \longrightarrow & Server \\
Server & \longrightarrow & File & \longrightarrow & Peer \\
Peer & \longrightarrow & File_{id} & \longrightarrow & Server \\
Server & \longrightarrow & List\ of\ peers & \longrightarrow & Peer
\end{array}
$$

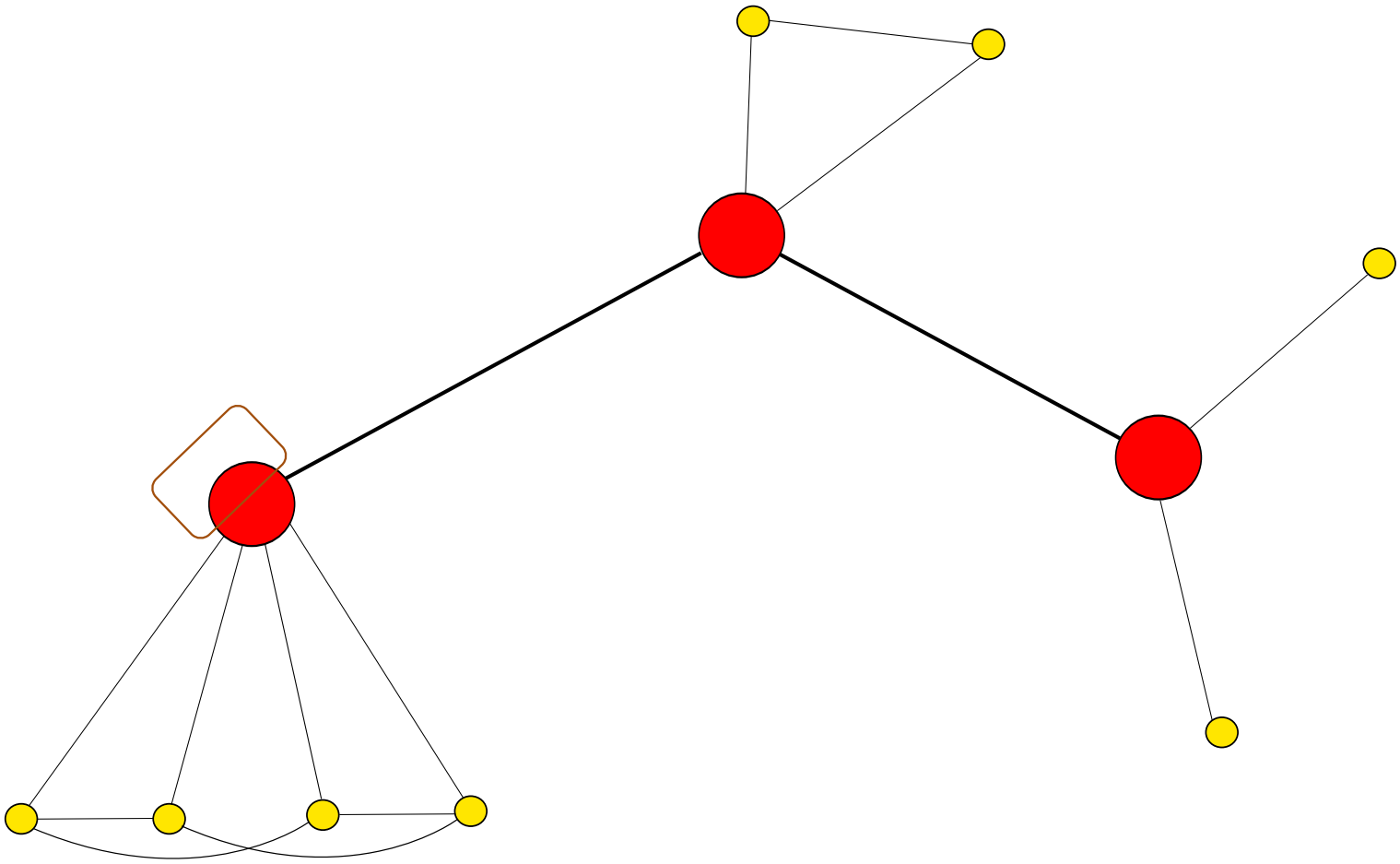
Method 1: measure on server



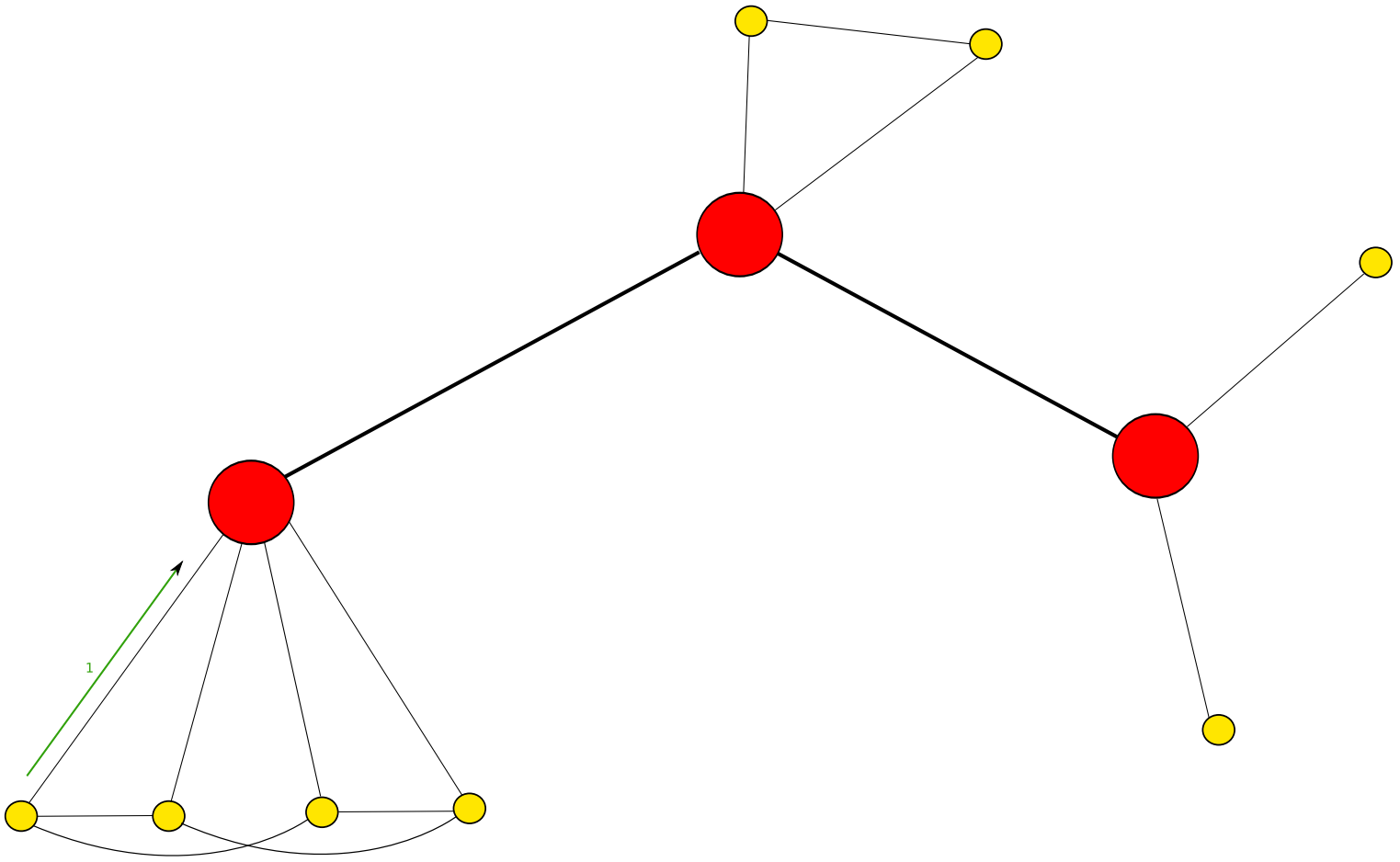
Method 1: measure on server

- 10 weeks of measurement
- 90 millions of peers
- 275 millions of file-ids
- Largest measurement of P2P

Method 1: measure on server

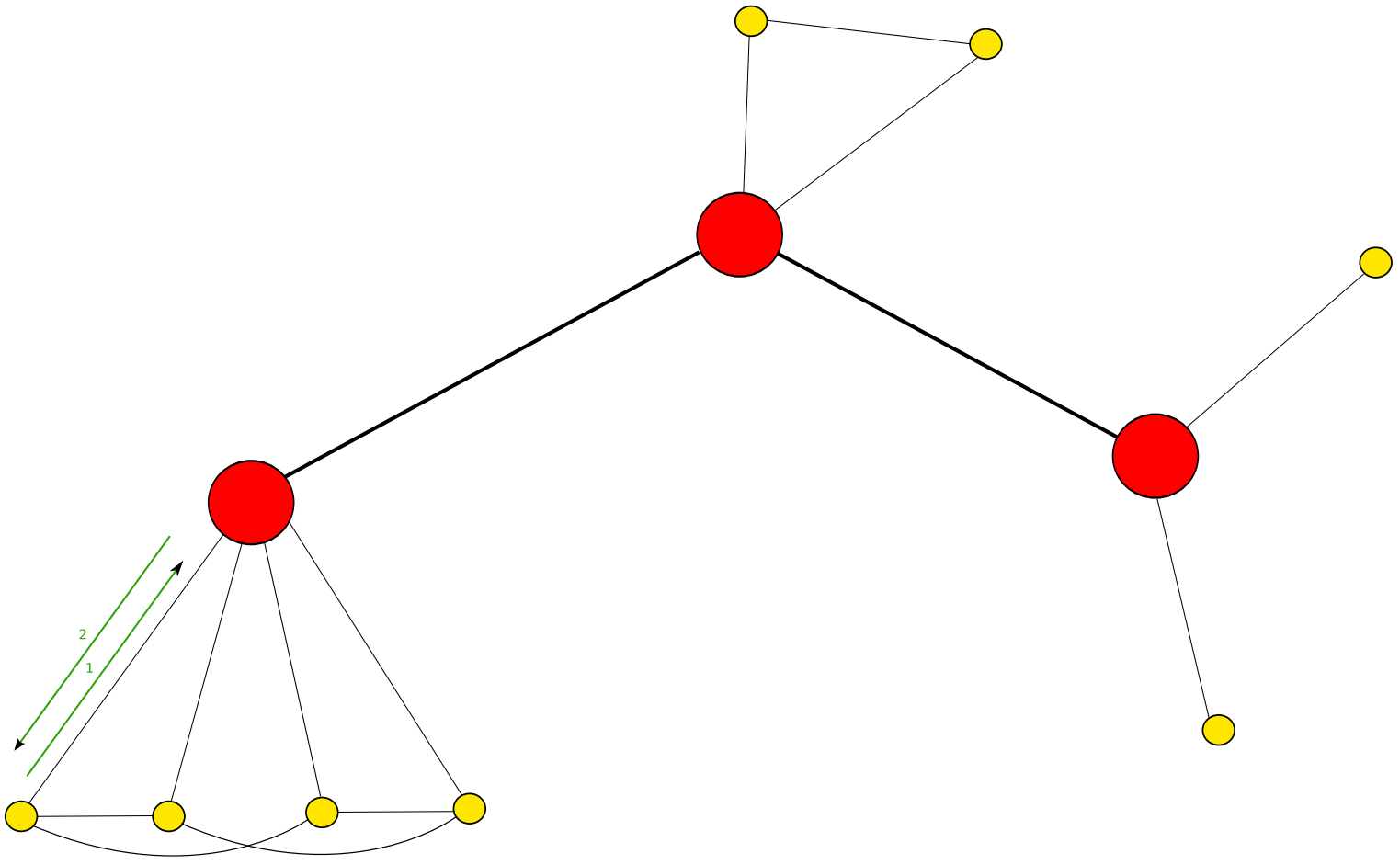


Method 1: measure on server

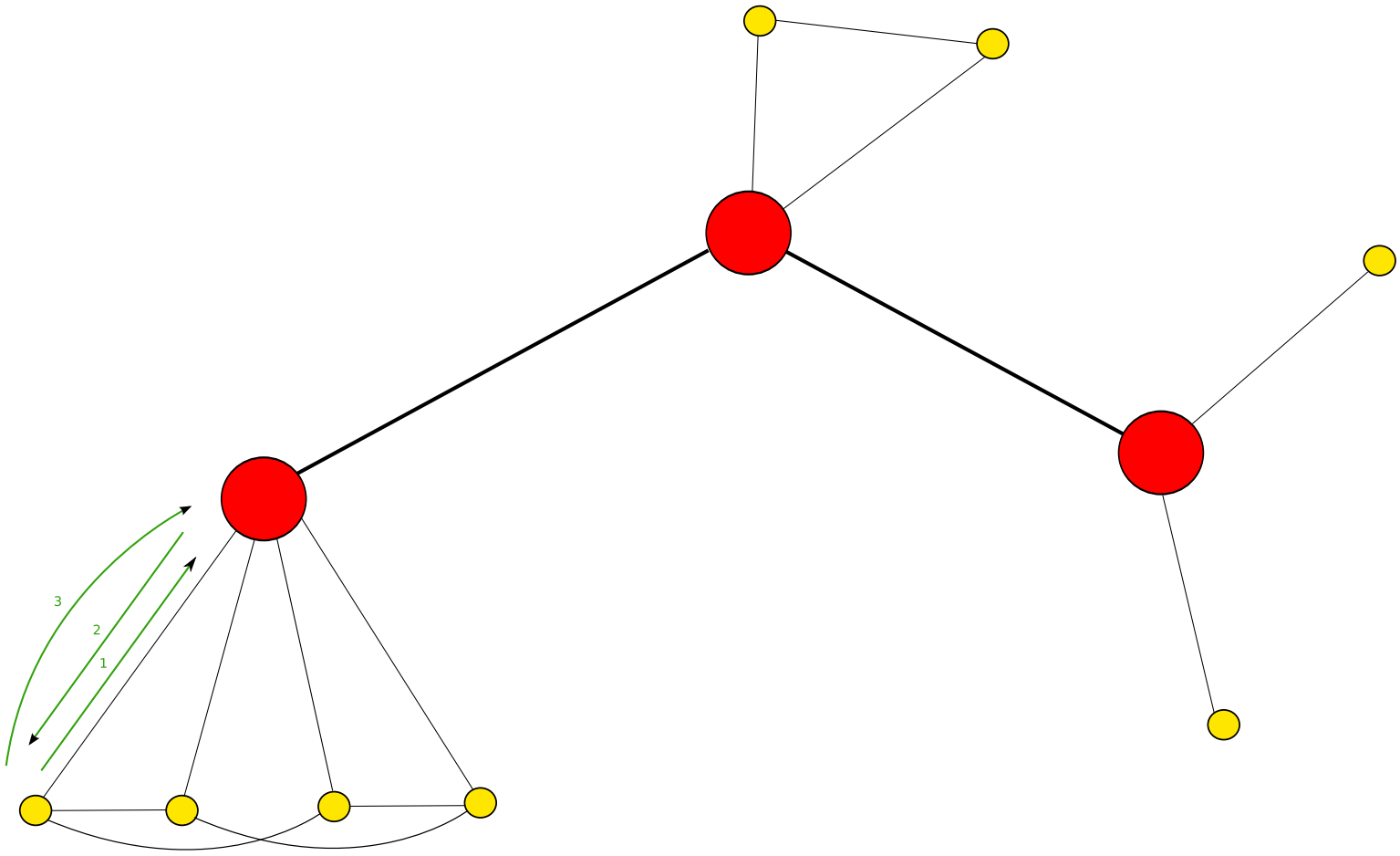


2. Measurement

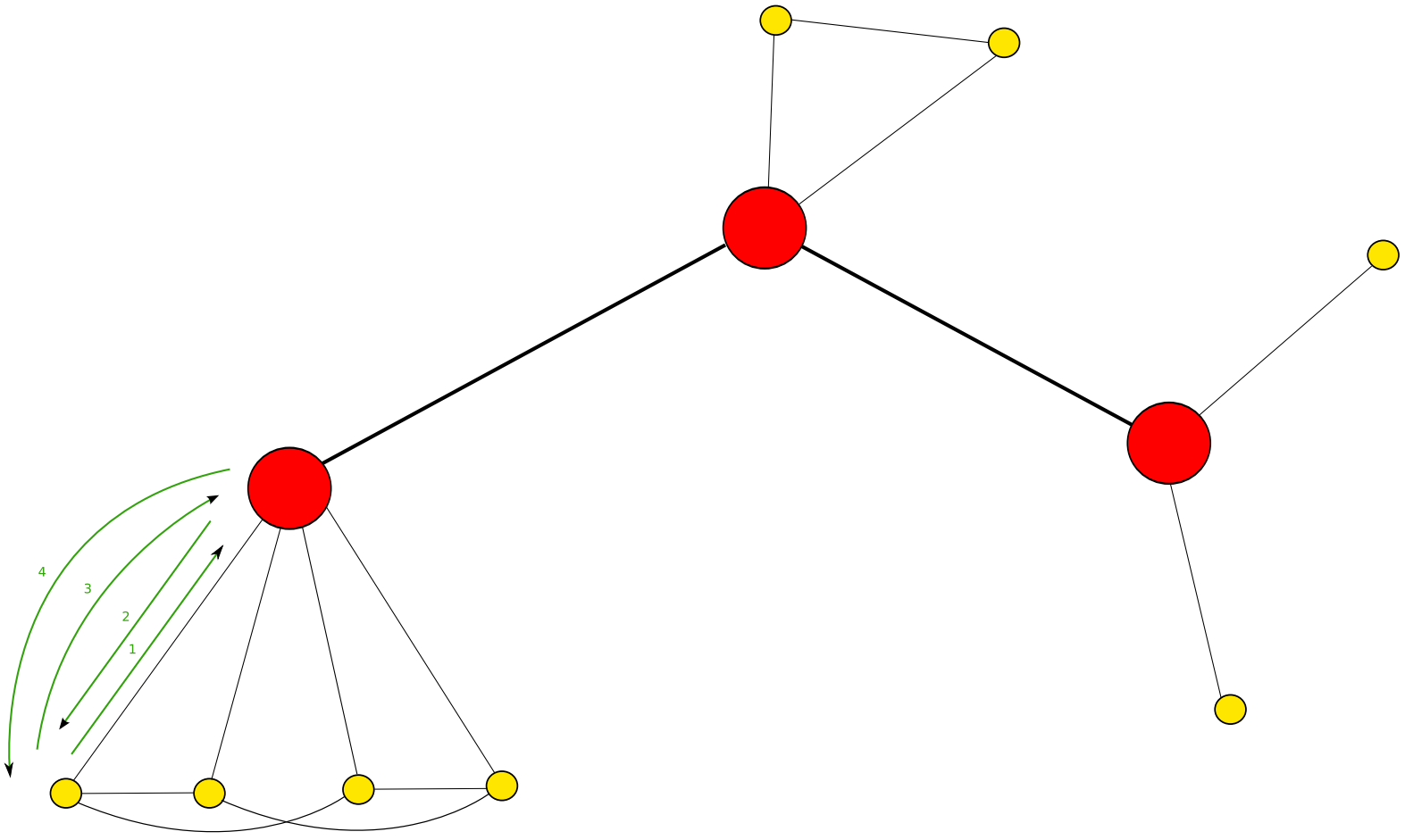
Method 1: measure on server



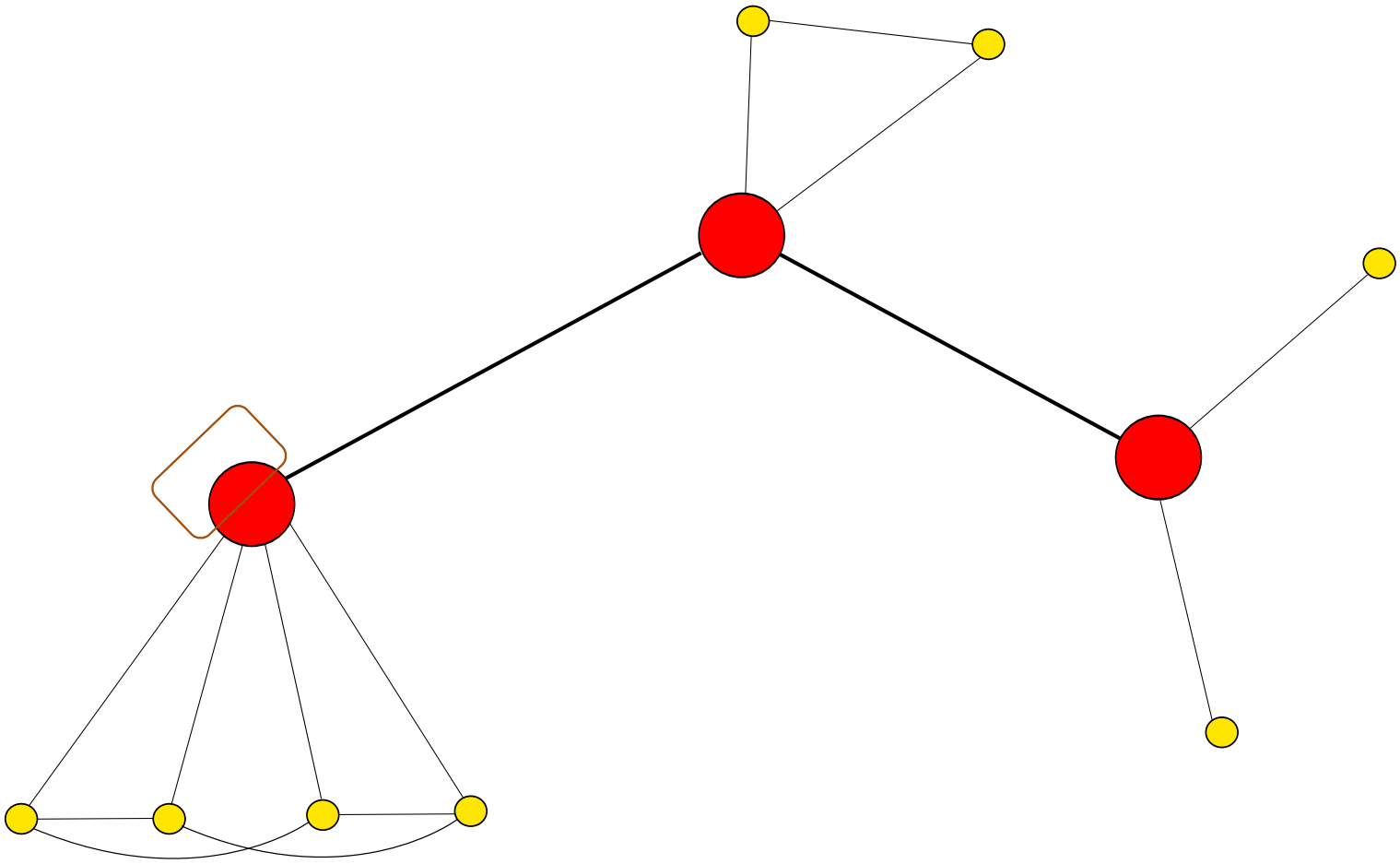
Method 1: measure on server



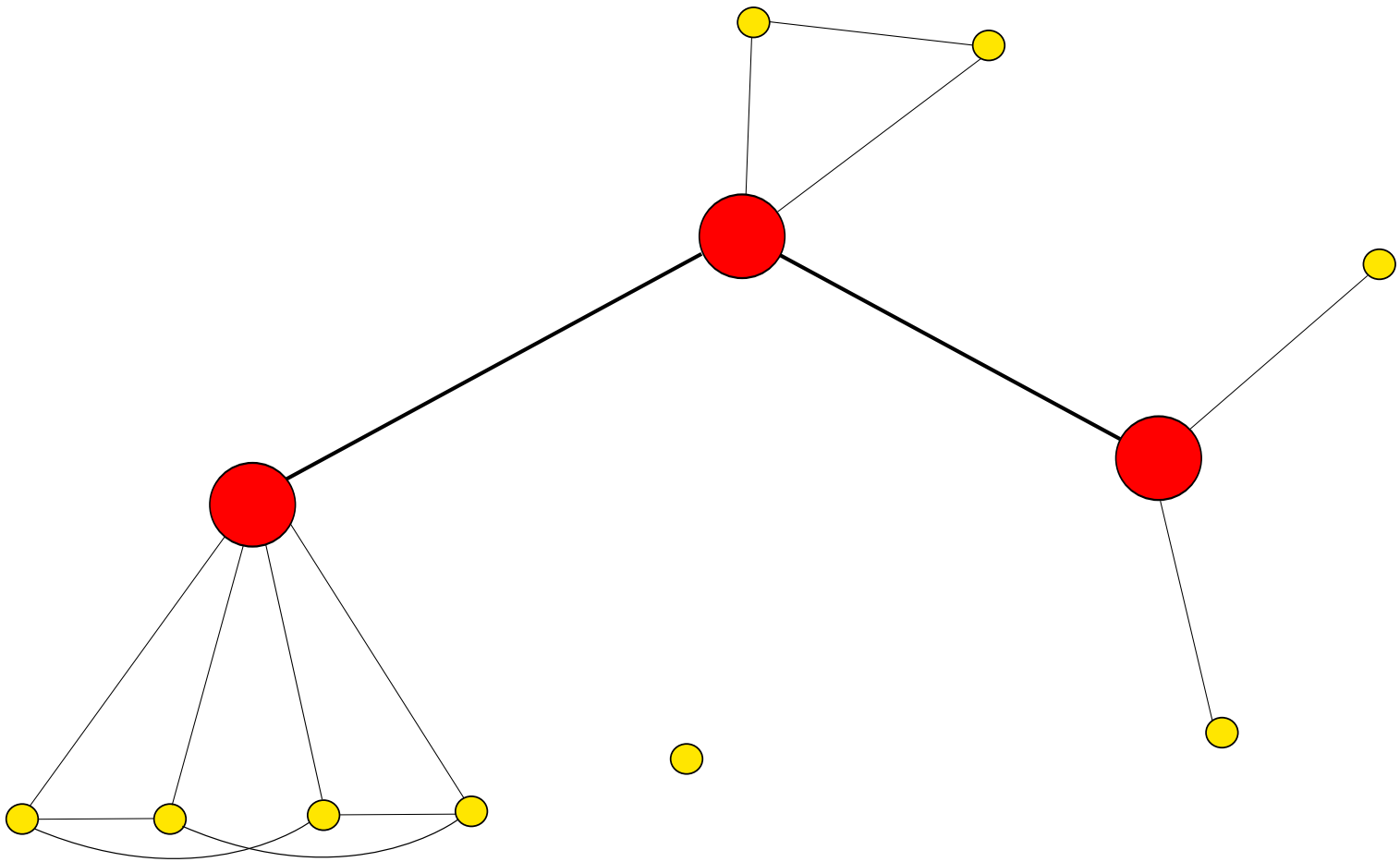
Method 1: measure on server



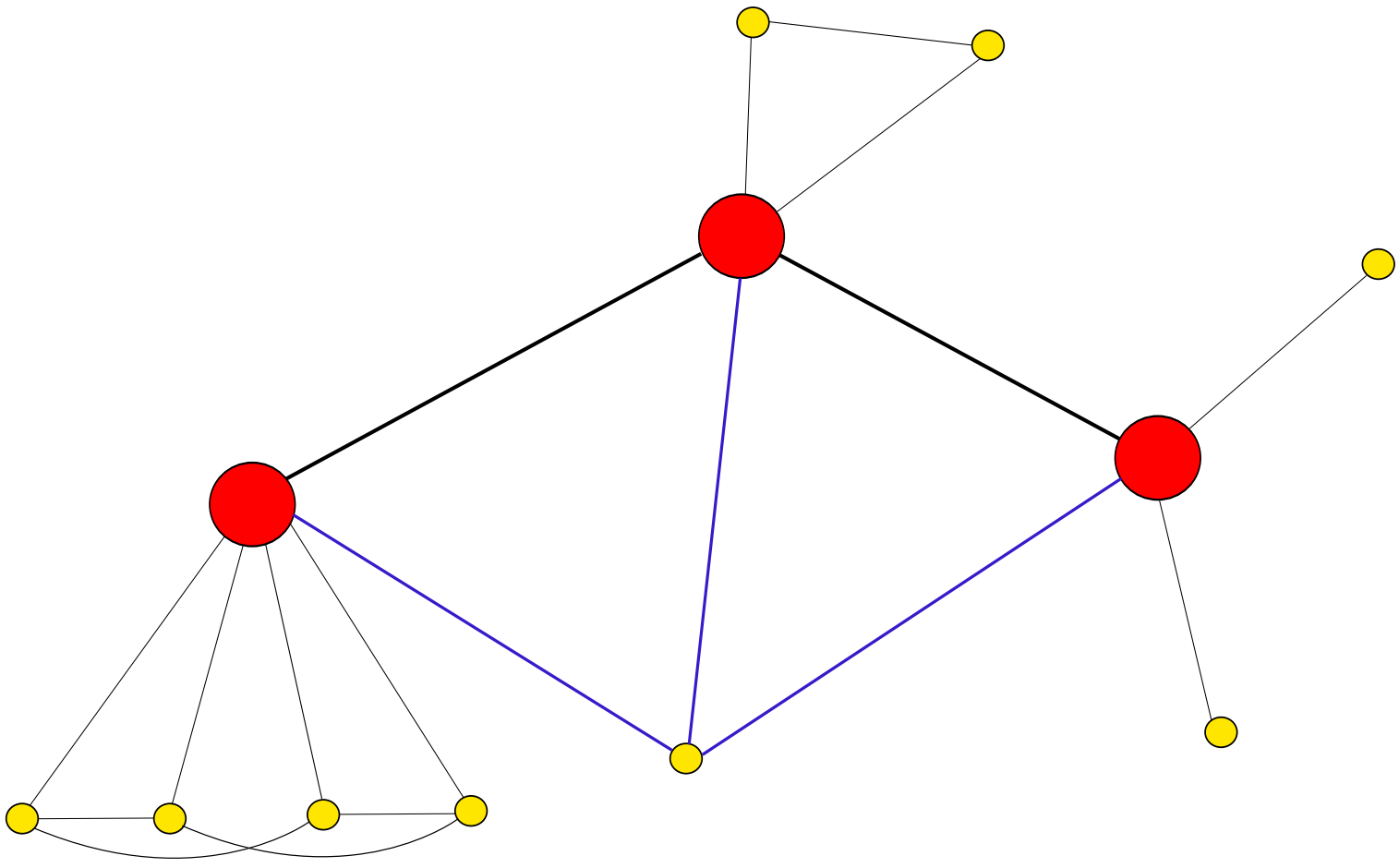
Method 1: measure on server



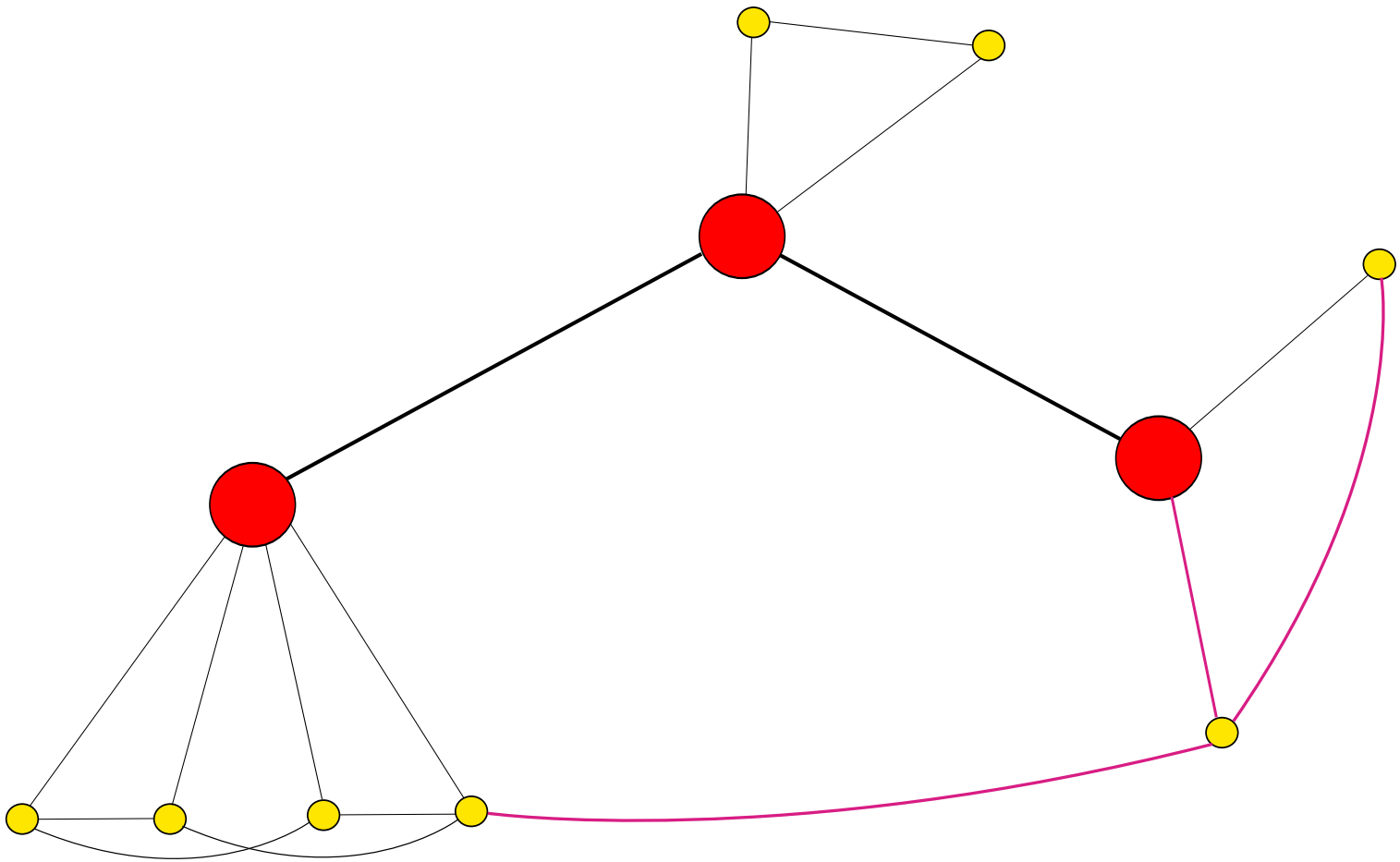
Method 2: measure by clients sending queries



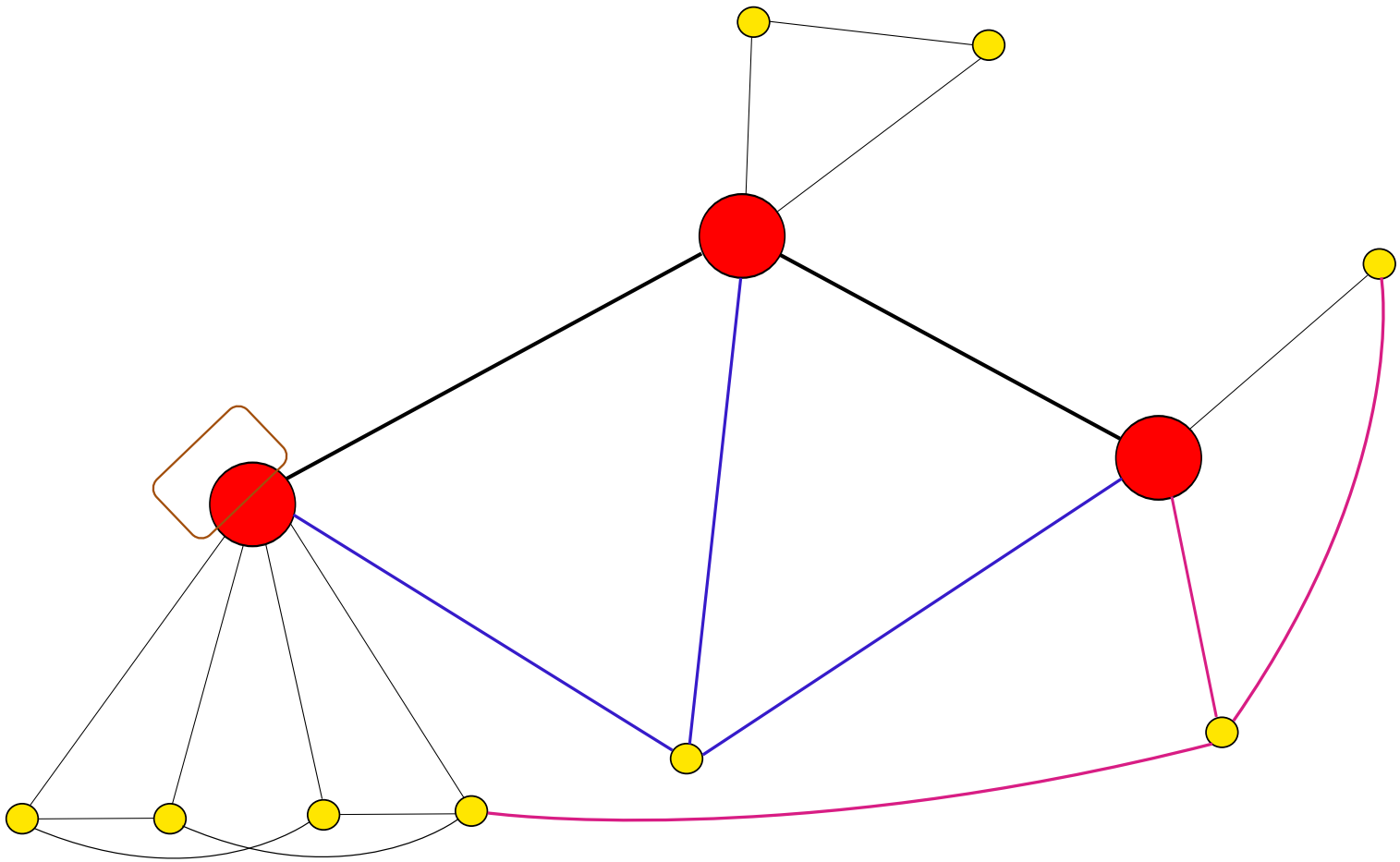
Method 2: measure by clients sending queries



Method 3: honey pot



2. Measurement



3. Data Processing

- Raw data:

$$
\begin{array}{ll}
T_1 & P \quad Q \\
T_2 & F_1\ F_2\ F_3 \longrightarrow P \\
T_3 & P \longrightarrow F_1\ F_3 \\
T_4 & F_1 : P_1\ P_{15} \longrightarrow P \\
 & F_3 : P_1\ P_3\ P_7 \longrightarrow P
\end{array}
$$

3. Data Processing

- Raw data:

$$T_1 \quad P \quad Q$$

3. Data Processing

- Raw data:

$$\begin{array}{cccc} T_1 & P & Q & \\ T_2 & F_1 & F_2 & F_3 \longrightarrow P \end{array}$$

3. Data Processing

- Raw data:

$$\begin{array}{ll} T_1 & P \quad Q \\ T_2 & F_1 \; F_2 \; F_3 \longrightarrow P \\ T_3 & P \longrightarrow F_1 \; F_3 \end{array}$$

3. Data Processing

- Raw data:

$$
\begin{array}{ll}
T_1 & P \quad Q \\
T_2 & F_1\ F_2\ F_3 \longrightarrow P \\
T_3 & P \longrightarrow F_1\ F_3 \\
T_4 & F_1 : P_1\ P_{15} \longrightarrow P
\end{array}
$$

3. Data Processing

- Raw data:

$$
\begin{array}{ll}
T_1 & P \quad Q \\
T_2 & F_1\ F_2\ F_3 \longrightarrow P \\
T_3 & P \longrightarrow F_1\ F_3 \\
T_4 & F_1 : P_1\ P_{15} \longrightarrow P \\
 & F_3 : P_1\ P_3\ P_7 \longrightarrow P
\end{array}
$$

3. Data Processing

- Web interface

3. Data Processing

- Web interface

  For each file id:

  - file names
  - number of providers
  - first provider (date and peer id)
  - main provider (list of peers id)
  - content rating
  - fake

3. Data Processing

- Web interface

  For each peer:

  - date of arrival
  - number of files shared
  - sent requests
  - provided files

3. Data Processing

- Anonymisation

3. Data Processing

- Anonymisation

  Strings $\overset{normalisation}{\longrightarrow}$ Normalised strings $\overset{cutting}{\longrightarrow}$ Words $\overset{computation}{\longrightarrow}$ Word frequency:

  ≤ 100 anonymisation

  ≥ 100 stored data

3. Data Processing

- $\geq$ 100 stored data: examples

3. Data Processing

- $\geq$ 100 stored data: examples

  Top "interesting" words (in file names and queries):
  you, *love*, *sex*, *xxx*

3. Data Processing

- $\geq$ 100 stored data: examples

  Top "interesting" words (in file names and queries):
  you, *love*, *sex*, *xxx*

  Top 15 words inqueries (not in file names):
  girl, *girls*, *boy*, *boys*, *child*, *children*, *playboy*, *pedo*, *attack*, *fille*, *enfants*, *boyz*, *enfant*, *incest*, *preteen*

4. Future

- More measurements (other methods)
- Content Rating and Fake Detection System
- Keywords analysis
- Better knowledge of paedophile activities