

Network analysis using Hadoop en Neo4j

Friso van Vollenhoven
@fzk

fvanvollenhoven@xebia.com

Why networks matter?



Kris Geusebroek
@krisgeus

 Follow



Booked ticket for Berlin buzzwords 4 and
5th June.

 Reply  Retweet  Favorite

3:05 PM - 7 Mar 12 via TweetDeck · Embed this Tweet

Reply to [@krisgeus](#)



RT: @krisgeus...

RT: @krisgeus...

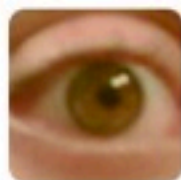
follows

@fzk



RT

RT: @krisgeus...



Tiziano Perrucci

@tizianoperrucci

 Follow



Cool, just got my **#GOTOAms** ticket. Thanks a lot to **@10gen** and **@fzk @agemooij** for organizing **#nlhug**.

 Reply  Retweet  Favorite

5:40 PM - 22 May 12 via web · Embed this Tweet

Reply to @tizianoperrucci @10gen @fzk @agemooij

Tiziano Perrucci
@tizianoperrucci

Cool, just got my #GOTOAms ticket. Thanks a lot to @10gen and @fzk @agemooij for organizing #nlhug.

3:40 PM - 22 May 12 via web - Embed this Tweet

Reply to @tizianoperrucci @10gen @fzk @agemooij

Twitter © 2012 Twitter About Help

goes to ↴

Netherlands Hadoop User Group

Home Members Sponsors Photos Pages Discussions More Group tools My profile



Amsterdam, Netherlands
Founded Jan 13, 2012

[Read more about us...](#)

Members **107**
Group reviews **2**

Welcome!

[+ SCHEDULE A NEW MEETUP](#)

Upcoming 1 Suggested 0 Past Calendar

Hadoop experiences at eBuddy and Enterprise ready Hadoop at Greenplum

Incentro Nootdorp **Thu Jun 14**
Gildeweg 5b, Nootdorp (map) **5:15 PM**

What's new

NEW MEMBER
Maarten Plesman joined 4 minutes ago

NEW DISCUSSION
Johan van der Kooij started next Hadoop meetup Thursday, June 14 Friday at 4:35 PM

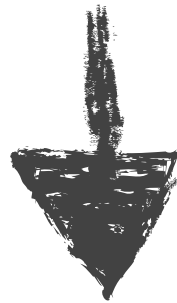
NEW RSVP

See any meaningful patterns?



A toy problem...

```
TABLE_DUMP2 | 1332345590 | B | 195.66.224.97 | 1299 | 1.11.64.0/21 |  
1299 6461 9318 38091 | EGP | 195.66.224.97 | 0 | 0 | | NAG | |
```



TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21
TABLE_DUMP2	1332345590	B	195.66.224.97	1299	1.11.64.0/21

raw data

transform

nodes.txt
+
edges.txt

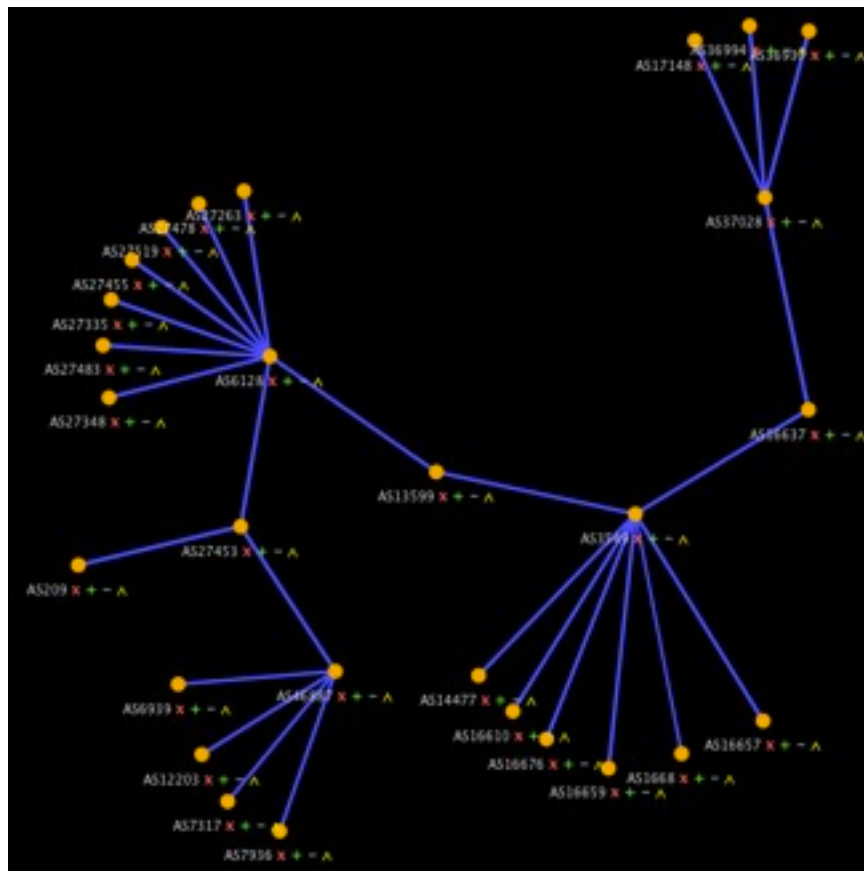
enrich?

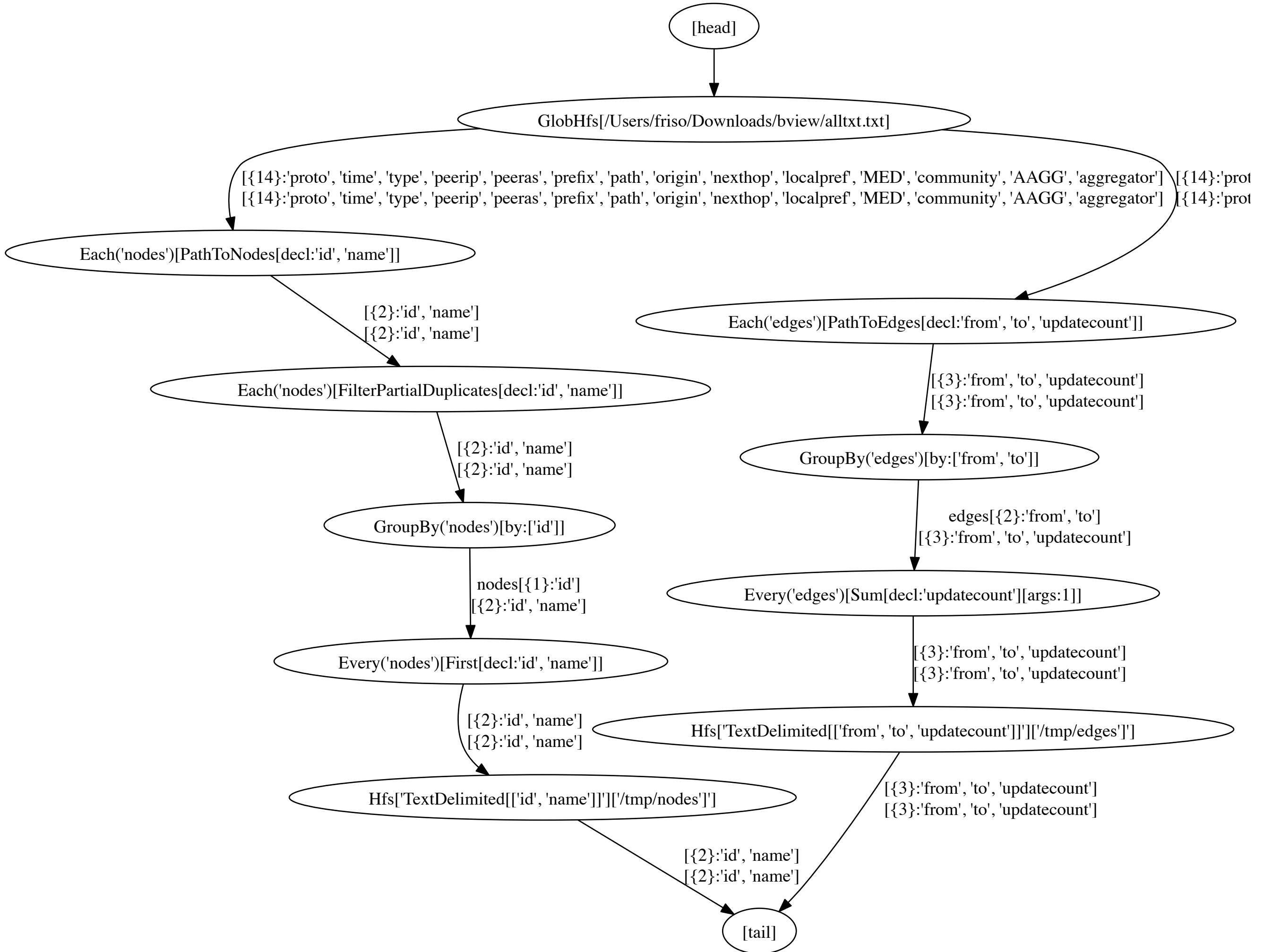
import

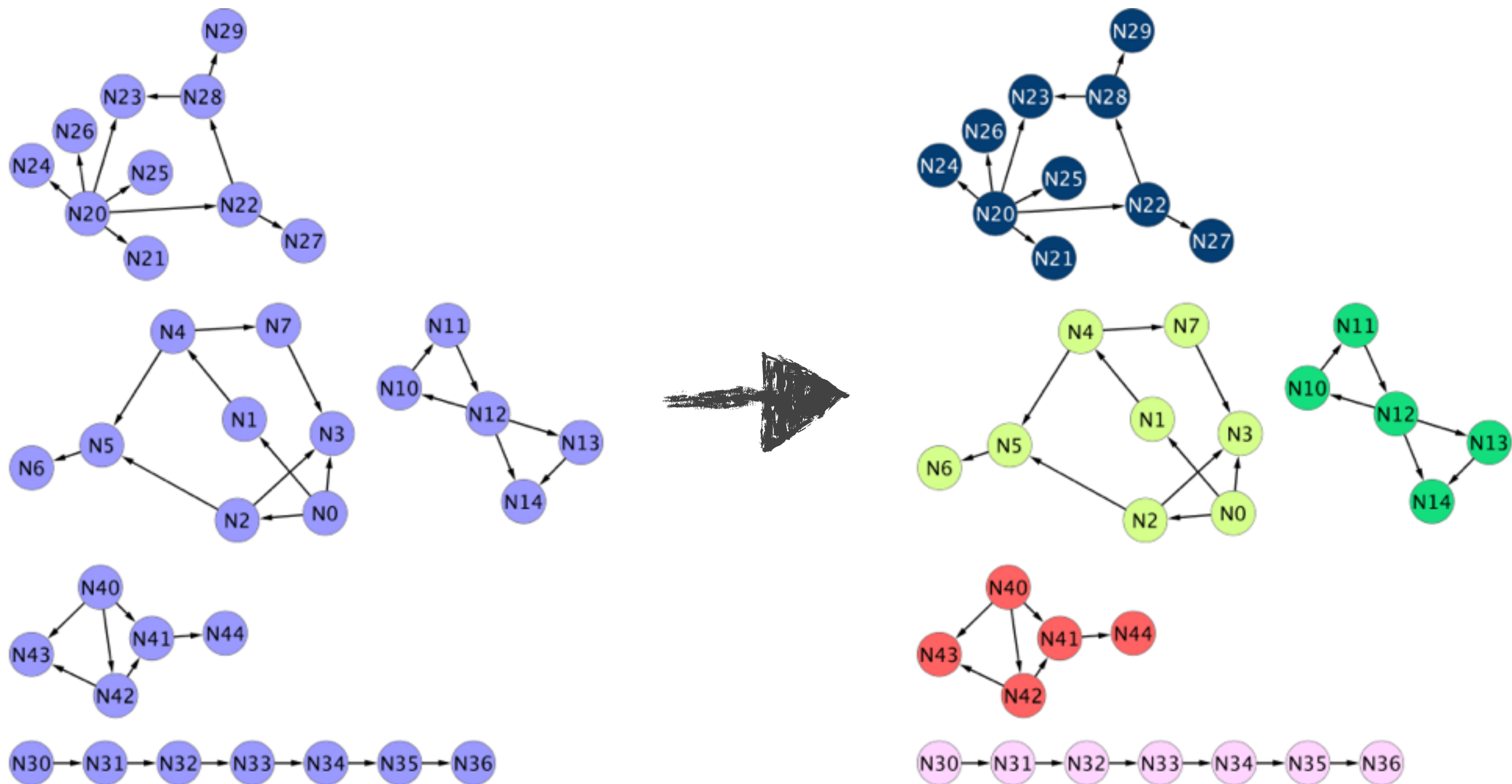
enriched-nodes.txt
+
enriched-edges.txt



query / interact







<http://bit.ly/lzWvcT> and <http://bit.ly/HHNNIb>

nodes.txt:

1	AS1	LVL1-1 - Level 3 Communications, Inc.
10	AS10	CSNET-EXT-AS - CSNET Coordination and Information Center (CSNET-CIC)
100	AS100	FMC-CTC - FMC Central Engineering Laboratories
1000	AS1000	GONET-ASN-17 - GONET
10000	AS10000	NCM Nagasaki Cable Media Inc.
10001	AS10001	MICSNET Mics Network Corporation
10002	AS10002	ICT IGAUENO CABLE TELEVISION CO.,LTD
10003	AS10003	OCT-NET Ogaki Cable Television Co.,Inc.
10004	AS10004	AS-PHOENIX-J JIN Office Service Inc.
10006	AS10006	SECOMTRUST SECOM Trust Systems Co.,Ltd.
10010	AS10010	TOKAI TOKAI Communications Corporation
10011	AS10011	ADVAN advanscope.inc
10012	AS10012	FUSION Fusion Communications Corp.

edges.txt:

1	21616	3
1	3705	3
1	2	3
2	3	1
3	4	2
3	11488	2
4	5	1
10	10	2
10	13227	2
12	12	1


```
public class SillyImporter {
    private static enum ConnectionTypes implements RelationshipType {
        FOLLOWS;
    }

    public static void main(String[] args) {
        BatchInserter database = new BatchInserterImpl("/Users/friso/Desktop/graph.db");
        BatchInserterIndexProvider provider = new LuceneBatchInserterIndexProvider(database);
        BatchInserterIndex index = provider.nodeIndex("allnodes", stringMap(
            "type", "fulltext",
            IndexManager.PROVIDER, "lucene"
        ));

        long fzkNodeId = database.createNode(map(
            new Object[] {"name", "fzk", "tweets", 25L} //node properties
        ));
        index.add(fzkNodeId, map(new Object[] {"name", "fzk"}));

        long krisgeusNodeId = database.createNode(map(
            new Object[] {"name", "krisgeus", "tweets", 100L} //node properties
        ));
        index.add(krisgeusNodeId, map(new Object[] {"name", "krisgeus"}));

        database.createRelationship(
            krisgeusNodeId, //from node
            fzkNodeId, //to node
            ConnectionTypes.FOLLOWS, //relationship type
            map(new Object[] {"retweets", 3})); //relationship properties

        index.flush();

        provider.shutdown();
        database.shutdown();
    }
}
```

30M nodes + 250M edges, < 30 minutes
(if your graph fits in memory)

Cypher graph query language

```
start a = node:allnodes('name:"fzk"')  
match p = a-[r]-b  
return p
```

All relationships from node with name=fzk

Cypher graph query language

```
start a = node:allnodes('name:"fzk"')
match p = a-[r]-b
where any(x in r.amounts where x > 500)
return p
```

All relationships from node with name=fzk, where any element in the array property amounts of the relationship is greater than 500

Cypher graph query language

```
start a = node:allnodes('name:"fzk"'),  
b = node:allnodes('name:"krisgeus"')  
match p = a-[:FOLLOWS*1..3]-b  
return p
```

All paths with a length between 1 and 3 (inclusive) between node with name=fzk and node with name=krisgeus with any relationship type=FOLLOWS

Cypher graph query language

```
start a = node:allnodes('name:"fzk"'),  
b = node:allnodes('name:"krisgeus"')  
match p = a<-[:FOLLOWS]-x-[:FOLLOWS]->b  
return x
```

All people who follow both fzk and krisgeus. Note the indication of direction in the relationship predicates.

JavaScript InfoVis Toolkit

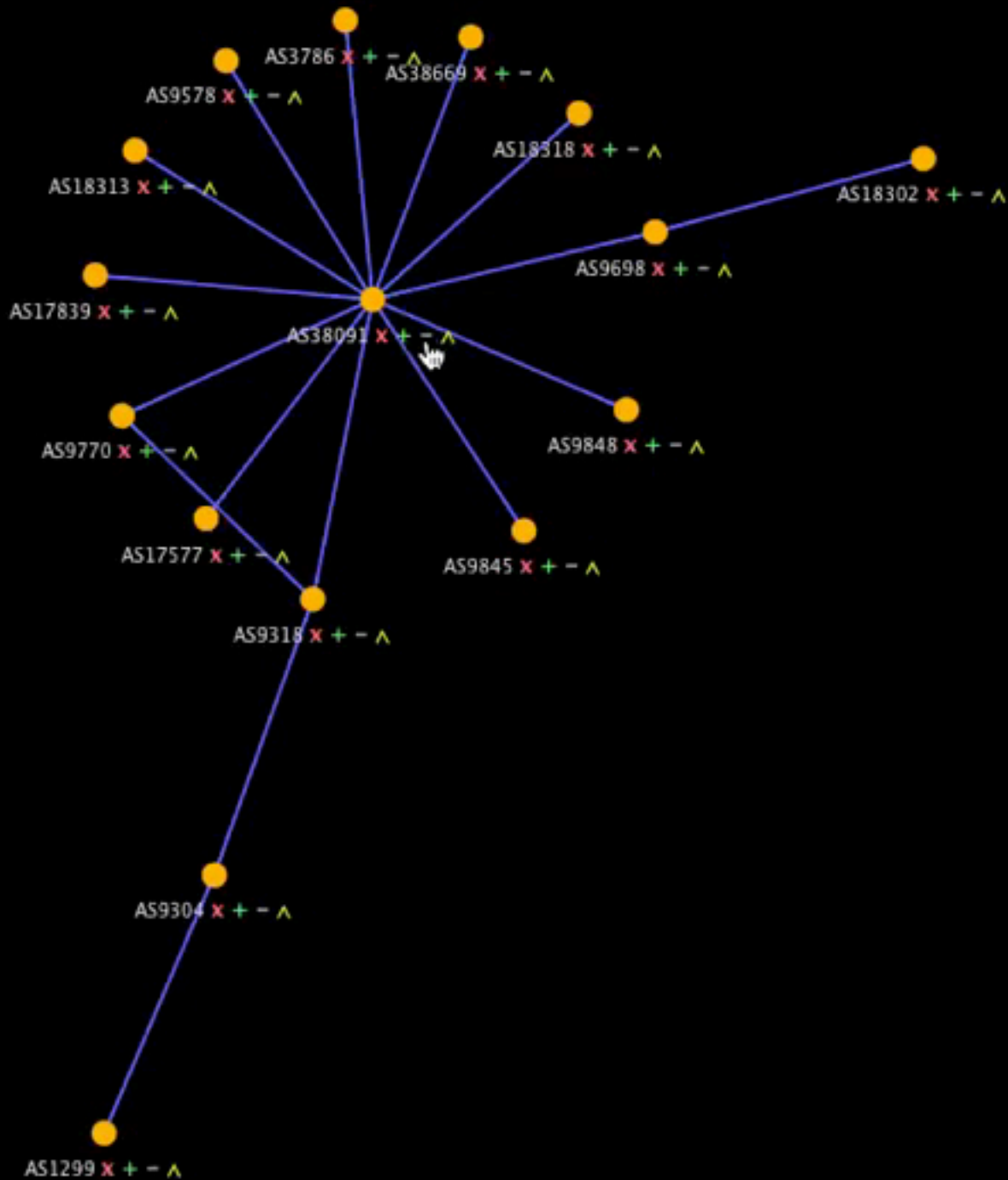
Create Interactive Data Visualizations for the Web

<http://thejit.org/>

```
start a = node(40518)
match p = a-[r]-(b)
where l=1
return p
```

ADD

REPLACE



Toy problem and viewer source code:
<https://github.com/friso/graphs>

Questions?

Friso van Vollenhoven
@fzk

fvanvollenhoven@xebia.com