

## Extracting and Learning Social Networks out of Multilingual News



Sonet 2008

20/09/2008

Bruno Pouliquen, Hristo Tanev, Martin Atkinson  
(& Web mining and intelligence team)

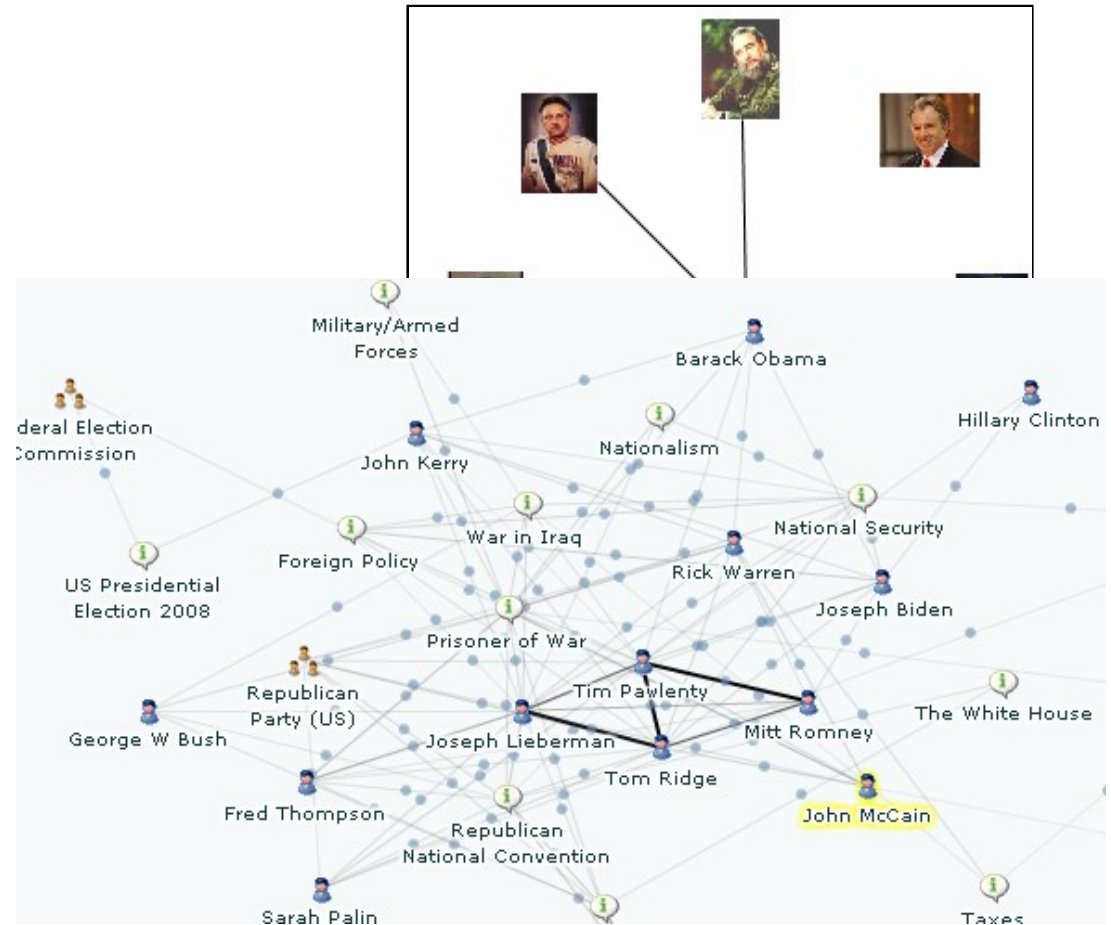
European Commission – Joint Research Centre (JRC)  
Institute for the Protection and Security of the Citizen (IPSC)  
Support to external Security Unit (SeS)

Web mining and intelligence

- Introduction
- Context
- Statistics-based social networks
  - Live-social networks
  - Long-term relationships between persons
- Linguistic patterns-based relationship
  - Syntax-based relationship
  - Quotation relationships
- Visualisation and Navigation
- Conclusion

- Manually building social network is not possible on large scale
- We need automatic tools
- Information is unstructured (texts)
  
- Corpus: EMM - multilingual news articles
- Highly multilingual, highly redundant
- We extract named entities in all articles:
  - Nodes: persons
  - Edges: two persons mentioned in same article (or cluster)
- Multilingual and cross-lingual

- Social networks based on user provided features
  - linkedIn
  - MySpace
  - Facebook
  - ...
- Presumably automatically built:
  - Connivence maps (en+fr)
  - Silobreaker (en)



→ We work on about 40 languages

- Statistics-based methods :
  - 😊 Good at recall
  - 😞 Bad at precision
  - 😊 Language independent
  - 😊 Cheap (Human input is minimum)
  - 😞 Cannot qualify the relationship
- Patterns-based methods :
  - 😊 Good at precision
  - 😞 Bad at recall
  - 😞 Language specific
  - 😞 Expensive (A lot of human input is required)
  - 😊 Can qualify the relationship

## Automatic developed systems for news aggregation and analysis

- **Gathering:** 50,000 news articles per day, in 40 languages, from 2000 news portals world-wide
- **Aggregation** of information (filtering, summarisation etc.)
- **Presentation** and visualisation (1.2 Million hits per day)

• **Tuesday, August 26, 2008**

### Suicide bomber targets recruits at Iraqi police station, 28 killed

- From different
- New name

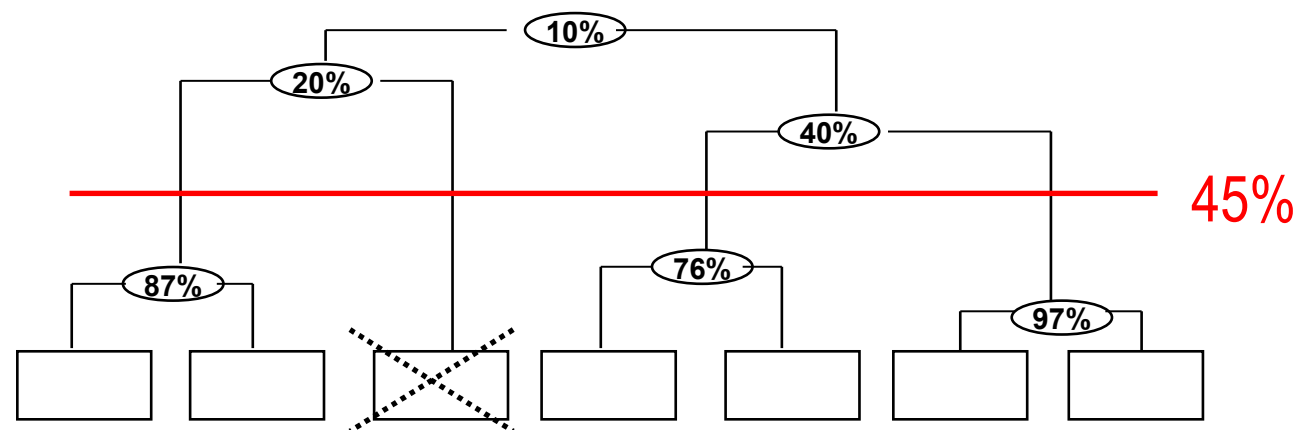
	Names	Key Titles and Phrases	External resources
ar	Nuri al-Maliki (da,sv)	iraqi prime minister (en - 1376)	
	Nouri al-Maliki (Eu,sv)	prime minister (en - 2826)	
	Al-Maliki (da,sv)	primeiro-ministro iraquiano (pt - 905)	
	Al Maliki (de,sv)	premier ministre irakien (fr - 673)	
	Nuri al Maliki (da,sl)	primer ministro iraquí (es - 504)	
	نوري المالكي (ar,fa)	iraakse premier (nl - 338)	
	Nouri al Maliki (da,ro)	ministerpräsident (de - 886)	
	Jawad al-Maliki (de,sv)	shi'ite prime minister (en - 242)	
	Nuri Kamal al-Maliki (de,sv)	premier (de,sl - 600)	
	Nouri Maliki (en,sv)	premier ministre (fr - 716)	
	Nuri el Maliki (de,tr)		
	Nuri Maliki (da,tr)		
	Nuri El Maliki (tr)		
	al Maliki (de,sl)		
	Nouri Al-Maliki (en,pt)		



Nouri al-Maliki (7)  
Saddam Hussein (7)  
George W. Bush (5)  
Falah Hassan (3)

- Input: Vectors consisting of keywords and country score
- Similarity measure: cosine
- Method: Bottom-up group average unsupervised clustering
- Build the binary hierarchical clustering tree (dendrogram)
  - Retain only “big” nodes in the tree with a high cohesion (empirically refined minimum intra-node similarity: 45%)
- Use the title of the cluster’s medoid as the cluster title
- For details, see Pouliquen et al. (CoLing 2004)

Keyness	Keyword
109.2478	jackson
41.5450	neverland
37.9347	santa
32.6105	molestation
24.5193	boy
24.4351	pop
20.6824	documentary
18.7973	accuser
13.5945	courthouse
11.1224	jury
10.4184	*us*
10.0838	ranch
9.6021	california
9.3905	verdict



en

death of former Prime Minister Rafik Hariri, blamed by many opposition

es

asesinato del exprimer ministro Rafic al-Hariri, que la oposición atribuyó

fr

l'assassinat de l'ex-dirigeant Rafic Hariri et le départ du chef de la diplom

nl

na de moord op oud-premier Rafiq al-Hariri gingen gisteren bijna een

de

libanesischen Regierungschef Rafik Hariri vor einem Monat wichtige B

sl

danjega libanonskega premiera Rafika Haririja. Libanonska opozicija si

et

möödumisele ekspeaminister Rafik al-Hariri surma põhjustanud pommip

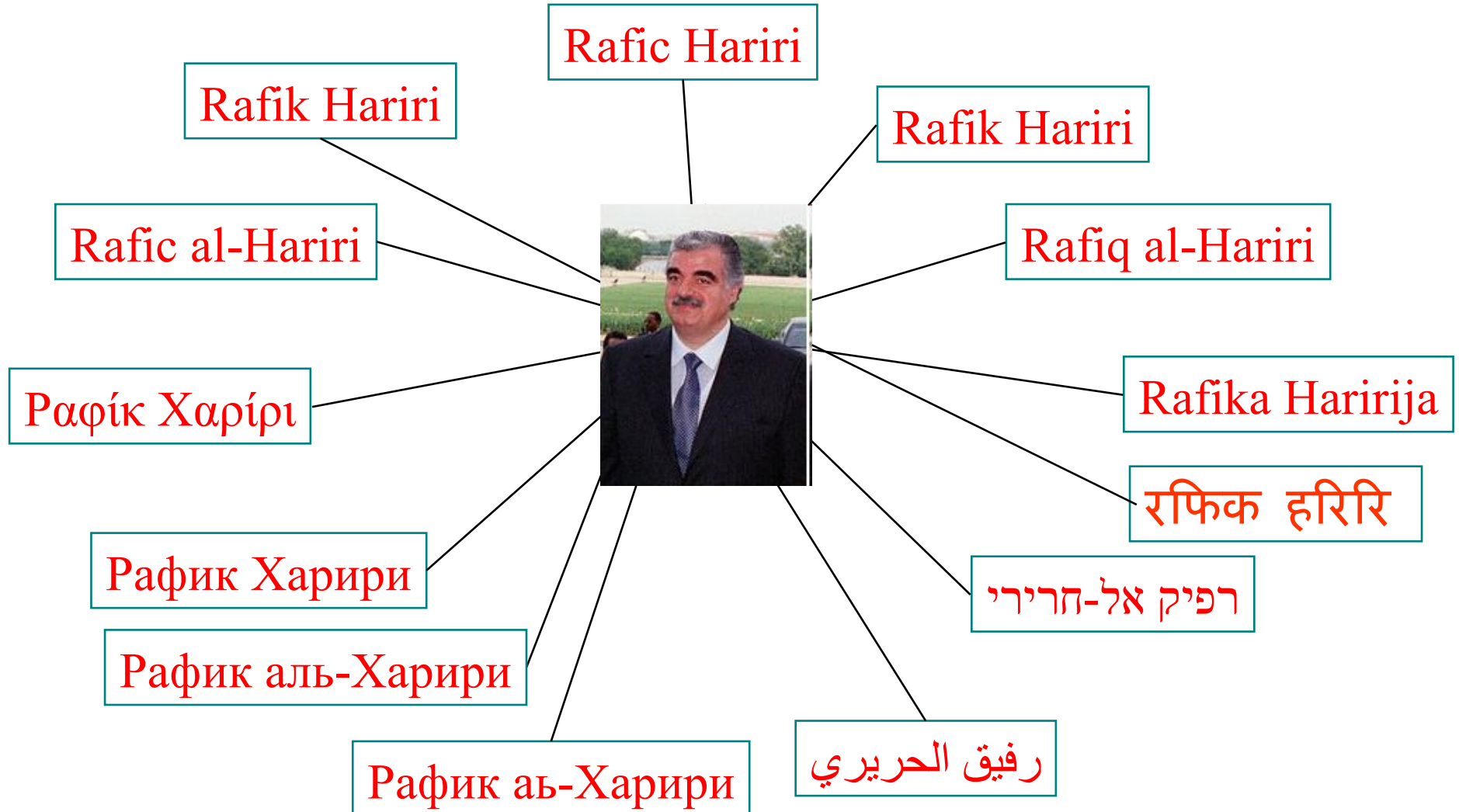
ar

اغتيال رئيس الوزراء السابق رفيق الحريري بأيد يهودية وما حدث سابقا

ru

БЫВШИЙ премьер-министр Ливана Рафик Харири, который

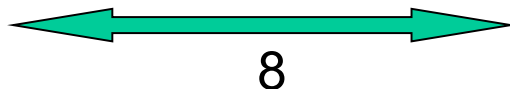




- Live-social networks
  - Who is with whom in the news now
- Long-term social networks
  - Who are the persons that appear together in a one year window

- Names are extracted from each incoming article
- Two names are linked if they appear in the same article
- Edges have the following attributes:
  - Undirected
  - Weight (number of articles where both names are mentioned)
  - Languages
  - Time
  - Link to text snippets
- Nodes
  - Person name
  - Picture
  - Frequency
  - Link to historically collected information

Lang	NewsPaper	Snippet
sl	vecер	glavnega osumljenca za umor <b>Aleksandra Litvinenka</b> v Londonu postavili pred
sl	vecер	v ponedeljek zavrnil izrocitev <b>Andreja Lugovoja</b> , da bi ga kot glavnega
tr	sabah	öldürülen eski KGB ajanı <b>Alexander Litvinenko</b> 'nun davası, İngiltere-Rusya
tr	sabah	cinayetin zanlısı olarak istediği <b>Andrei Lugovoy</b> 'u Rusya'nın iade etmemesi
en	dailytimesPK	suspected of killing Kremlin critic <b>Alexander Litvinenko</b> in London last year,
en	dailytimesPK	when British prosecutors alleged that <b>Andrei Lugovoi</b> used a rare radioactive
pt	DiariodeNoticias	assassinio do ex-oficial do KGB <b>Alexander Litvinenko</b> . A revelação foi feita
pt	DiariodeNoticias	acederia ao pedido de extradição de <b>Andrei Lugovoi</b> (outro ex-agente do KGB)
en	taipeitimes	Kremlin following its refusal to extradite <b>Andrei Lugovoi</b> , the former KGB....
en	taipeitimes	KGB agent suspected of murdering <b>Alexander Litvinenko</b> last November.
en	eirepost	Lugovoi over the murder of <b>Alexander Litvinenko</b> , describing the decision
en	eirepost	Russia's refusal to extradite <b>Andrei Lugovoi</b> over the murder of Alexander
sl	delo	in nekdanjega tajnega agenta KGB <b>Andreja Lugovoja</b> . London - Britanija in
sl	delo	in ostrega Putinovega kritika <b>Aleksandra Litvinenka</b> , ki je bil nekoc prav tako
en	rian	- Russia considers the <b>Alexander Litvinenko</b> case a purely criminal matter,
en	rian	Moscow has refused to extradite <b>Andrei Lugovoi</b> , a former Kremlin bodyguard,



## 2007-07-13T06:38+0200 . [UN nuke delegation arrives in Iran](#) [iranmania]

...to meet with Iran's top nuclear negotiator, **Ali Larijani**, later in the day, the report said. Accordiated Press (AP), Larijani and IAEA Chief **Mohammad ElBaradei** met last month in Vienna, Austria. Earli...

LONDON, July 12 (IranMania) - Iran's President Mahmoud Ahmadinejad said that the West should not expect his country to suspend uranium enrichment activities, the official Islamic Republic News Agency reported.

## 2007-07-13T06:37+0200 . [OIEA asegura haber logrado acuerdos Irán](#) [HoyDigital-DO]

...Internacional de Energía Atómica (OIEA), **Mohamad el Baradei**, hizo esta declaración al término de la (...) i, el asesor del principal negociador iraní **Ali Larijani**, que preside la parte iraní en las negociac...

TEHERAN, (EFE).- El jefe de la delegación del OIEA que visita Irán, Olli Heinonen, afirmó ayer que su equipo ha alcanzado un acuerdo con los dirigentes iraníes sobre "algunas cuestiones" en las negociaciones entre las dos partes sobre el caso nuclear iraní.

## 2007-07-13T02:13+0200 . [REGION: Iran, UN team hold talks on nuclear issues](#) [dailytimesPK]

... deputy to Iran's chief nuclear negotiator, **Ali Larijani**. President Mahmoud Ahmadinejad said on Wedn (...) unciil. The UN watchdog's Director General **Mohamed ElBaradei** has said Iran's transparency offer combi...

TEHRAN: Iranian nuclear officials and a visiting team from the UN nuclear watchdog held a second round of talks on Thursday to discuss ways to remove outstanding questions about Iran's disputed nuclear programme. Iran has offered to draw up an "action plan" to address Western suspicions that its nuclear programme is a front to obtain nuclear arms.

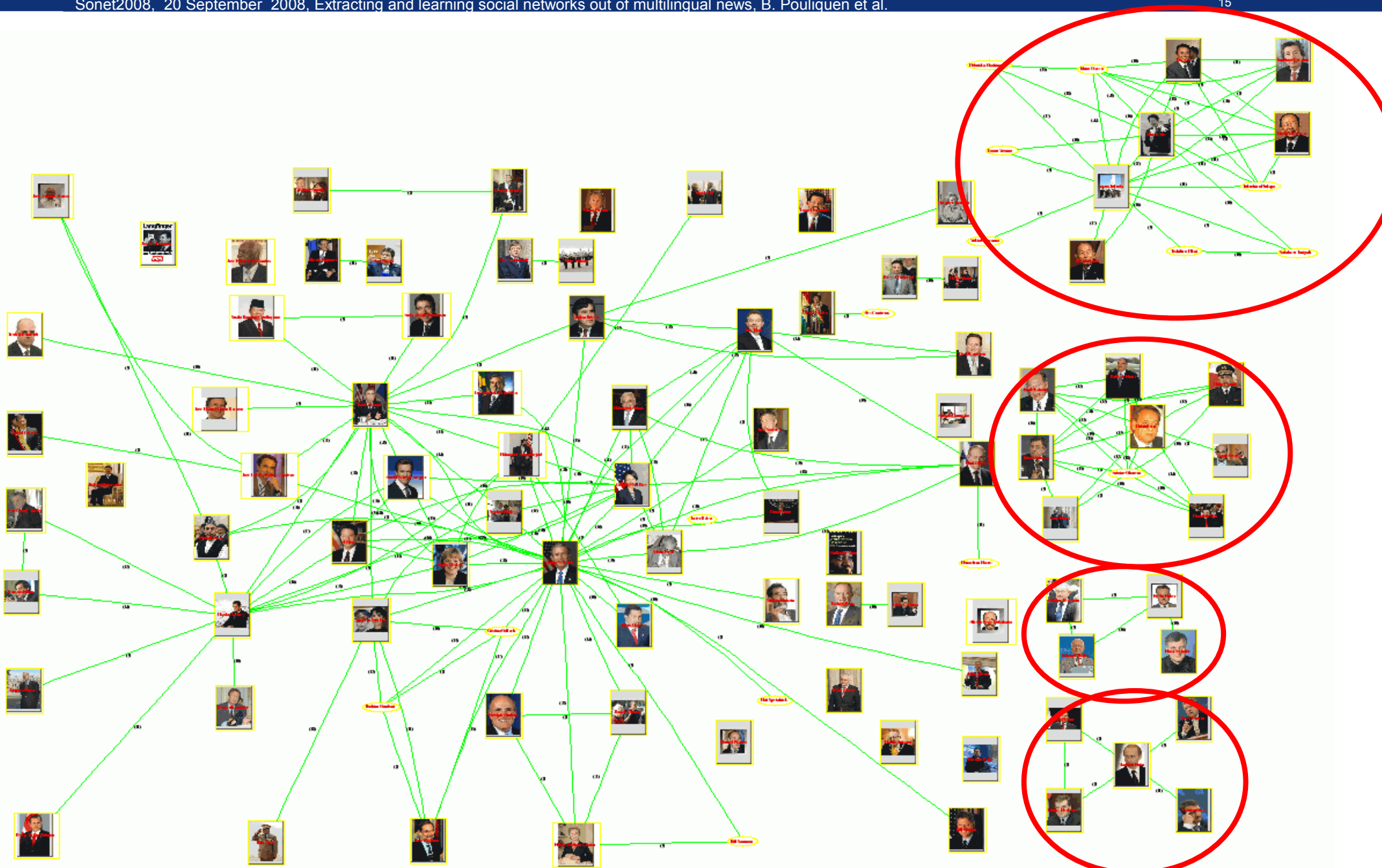
## 2007-07-13T01:31+0200 [نتائج بناءة بين ايران والوكالة الدولية للطاقة](#) [alrai]

... في ذلك بين كبير المفاوضين النوويين الإيرانيين **علي لاريجاني** ورئيس الوكالة الدولية للطاقة الذرية **محمد البرادعي** في الشهر الماضي في حينما يشافو سبعاة الاجراء...

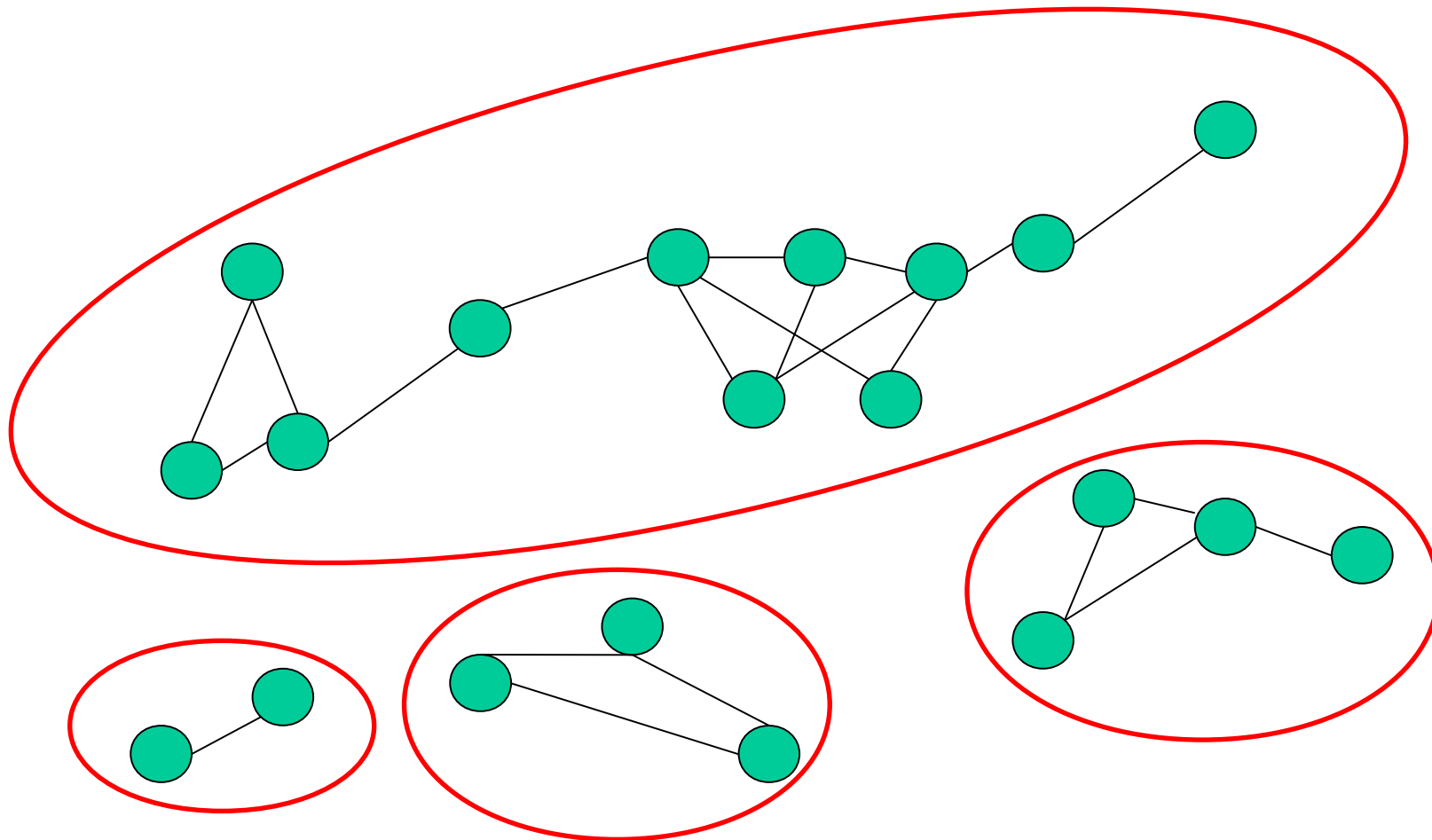
طهران - وكالات - اعلنت ايران امس انه تم التوصل الى نتائج جيدة بعد ثلاث جولات من المحادثات مع وفد من الوكالة الدولية للطاقة الذرية. واختتمت الجولة الثالثة من المحادثات حول البرنامج النووي الإيراني بين المسؤولين الإيرانيين ووفد الوكالة الدولية للطاقة الذرية برئاسة أولي هاننونين نائب مدير الوكالة الدولية للطاقة الذرية.

For a given time window (from midnight up to now)

- Compute all the links having minimum weight (min:4 → 50,000 links per day)
- Undirected graph
- (Non-connected) graph
  - isolated groups



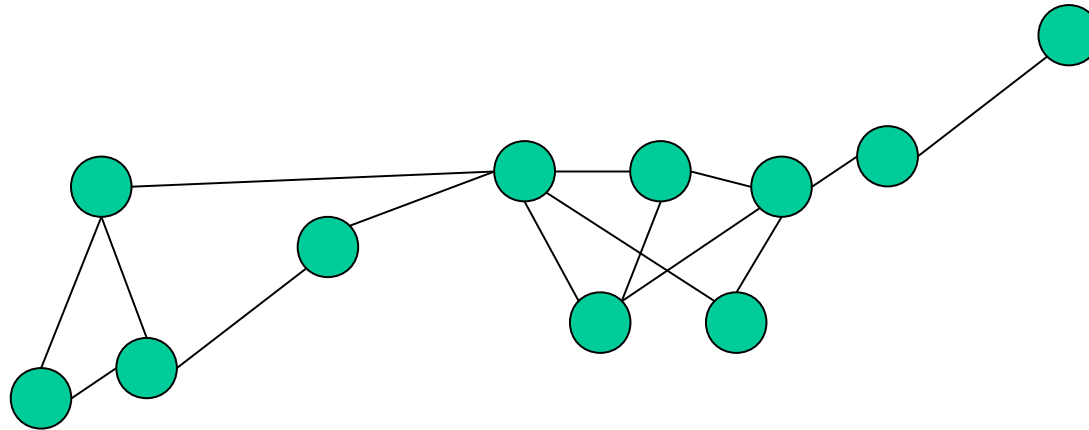
- Hard partitioning: all connex graphs



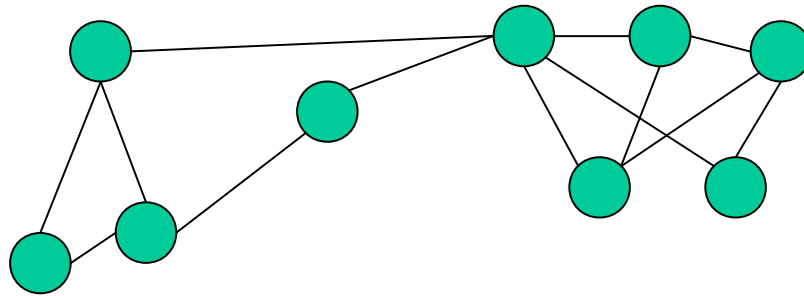


- Display the 12 biggest sub-graphs
- “shaving” big sub-graphs
- Algorithm

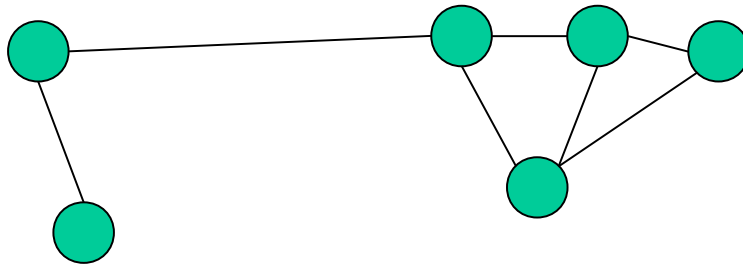
- Delete nodes having one link



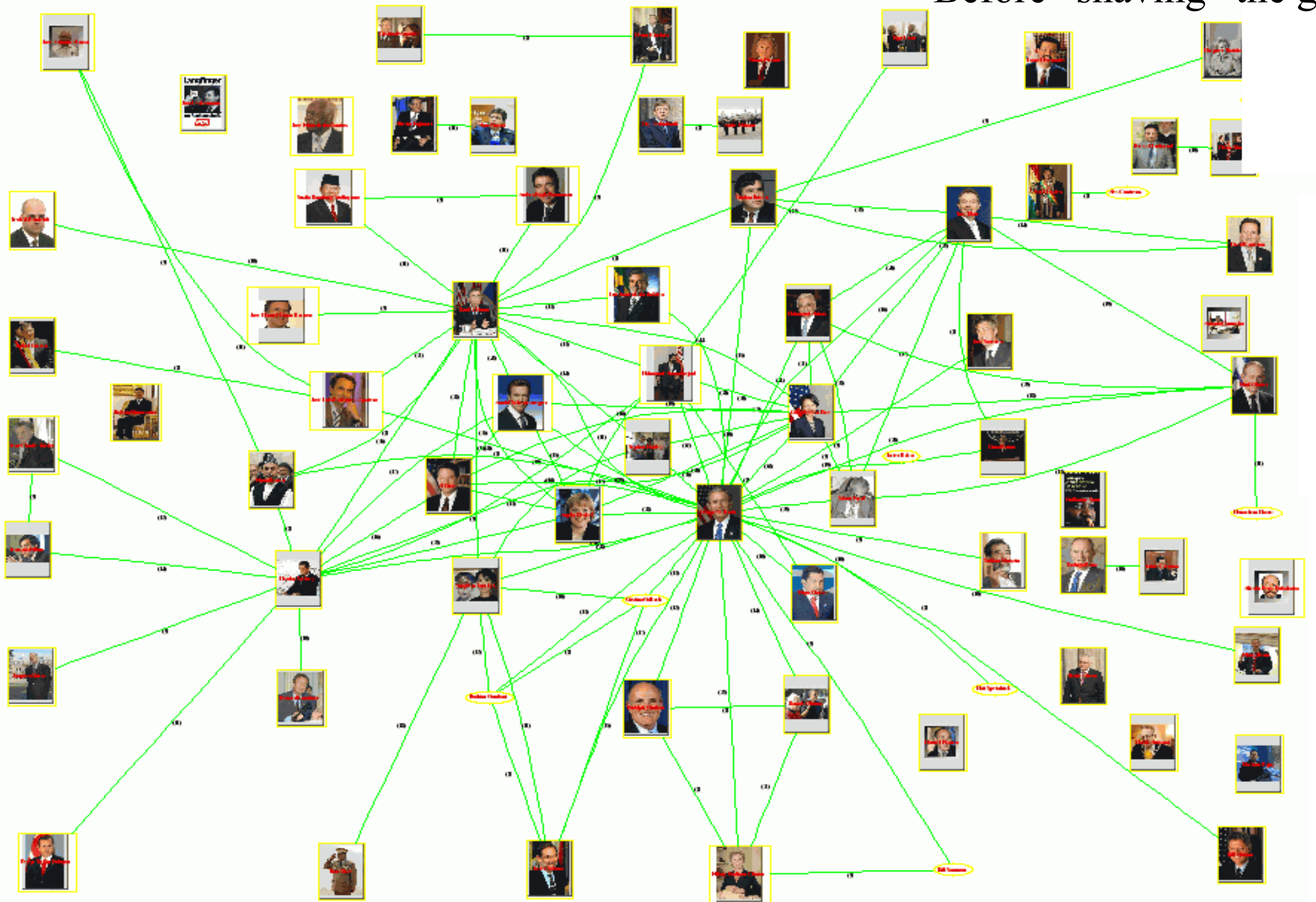
- Delete nodes having two links



- Stop when the number of links is less than the threshold



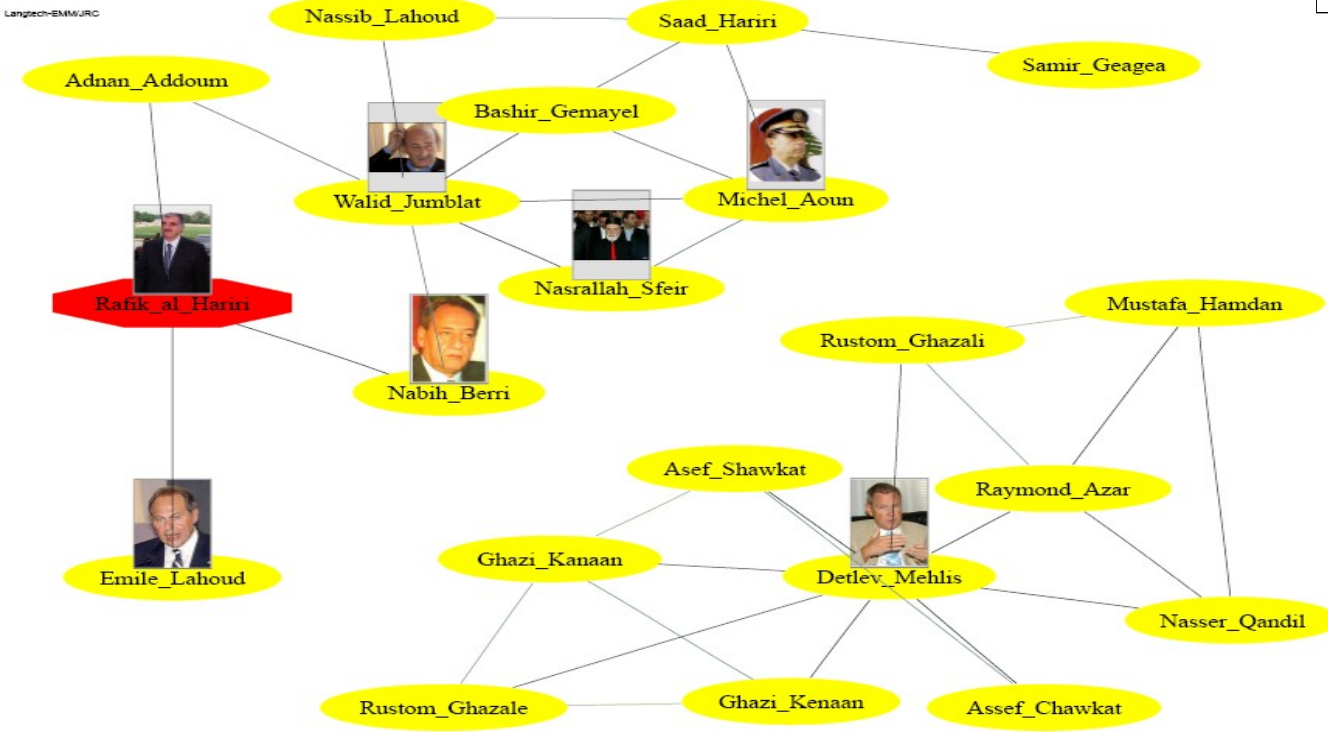
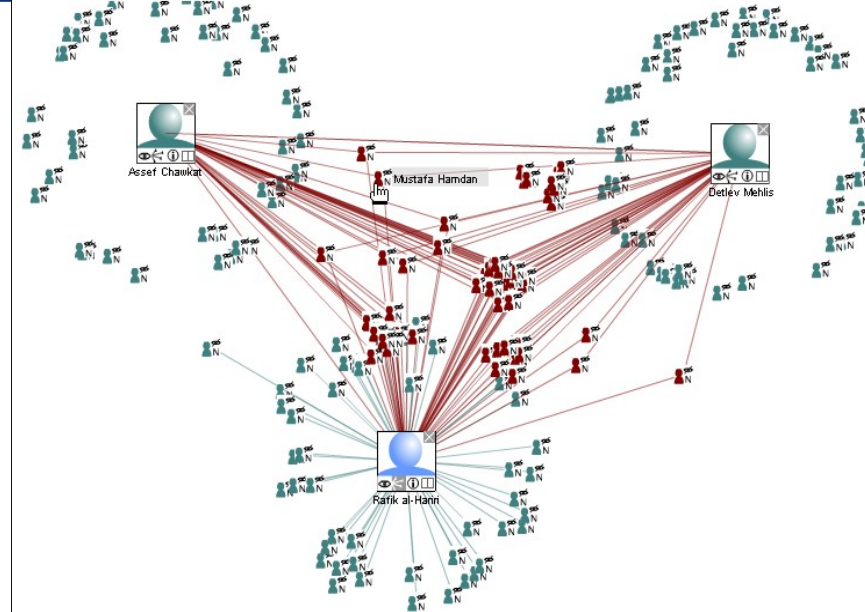
Before “shaving” the graph



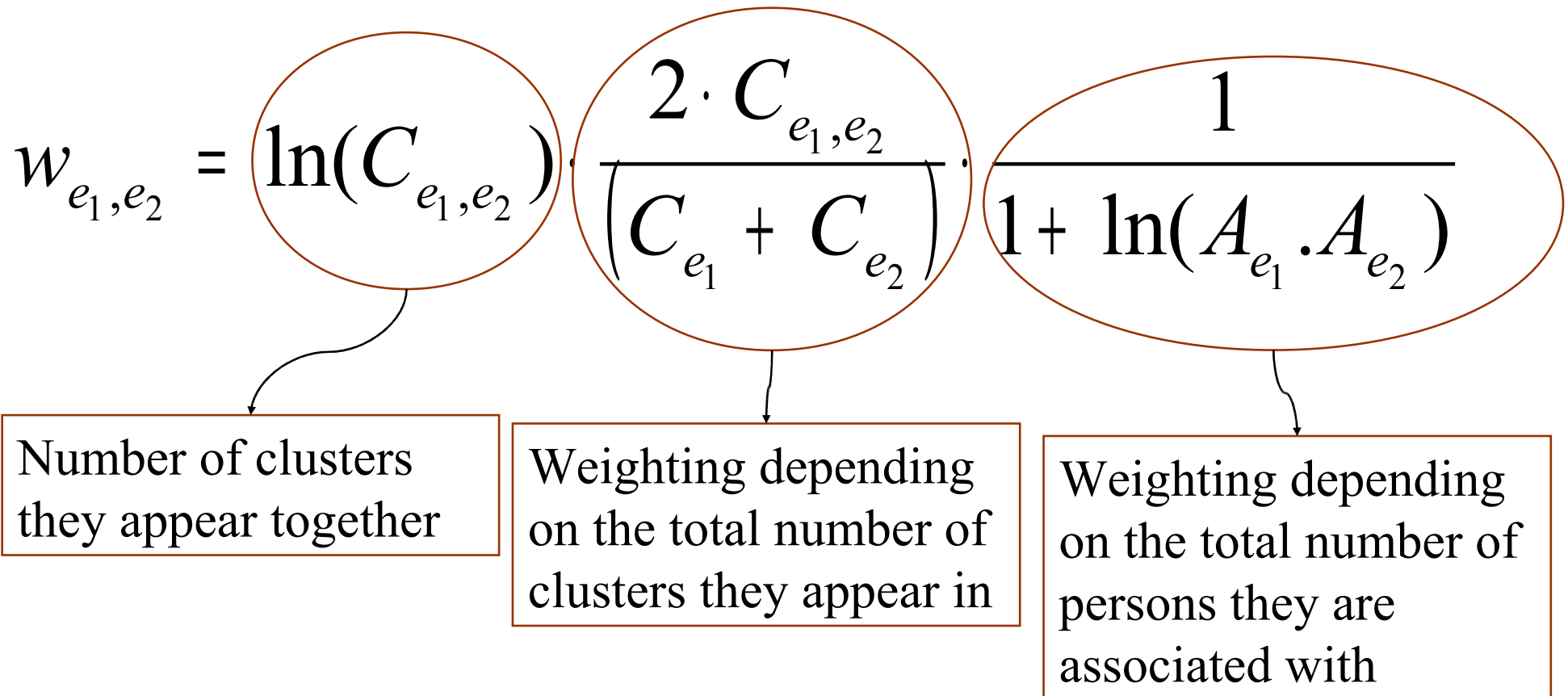


- All articles are clustered
- The cluster information is stored in our knowledge base
  - Articles, language, persons...
- We can easily query the persons that appear in same cluster

- We can extract person-person relationship
  - For a given person
  - For a given time period, language etc.
- Currently we display the social network around a given person





$$w_{e_1, e_2} = \ln(C_{e_1, e_2}) \cdot \frac{2 \cdot C_{e_1, e_2}}{(C_{e_1} + C_{e_2})} \cdot \frac{1}{1 + \ln(A_{e_1} \cdot A_{e_2})}$$


Number of clusters they appear together

Weighting depending on the total number of clusters they appear in









Weighting depending on the total number of persons they are associated with

*Javier Solana*










*(European Union's current Foreign Secretary)*

## Without weighting

	European Union	O	264	(de..sl)	(19-JUL-04..17-MAR-05)
	George W. Bush	P	191	(de..sl)	(19-MAR-04..12-MAR-05)
	Jacques Chirac		140	(de..sl)	(21-MAR-04..15-MAR-05)
	United Nations	O	131	(de..nl)	(10-JUL-04..15-MAR-05)
	Yasser Arafat		130	(de..sl)	(22-MAR-04..04-MAR-05)
	Kofi Annan		127	(de..sl)	(15-APR-04..16-MAR-05)
	Jose Manuel Durao Barroso		123	(de..sl)	(18-JUN-04..10-MAR-05)
	Ariel Sharon		115	(de..sl)	(29-MAR-04..17-MAR-05)

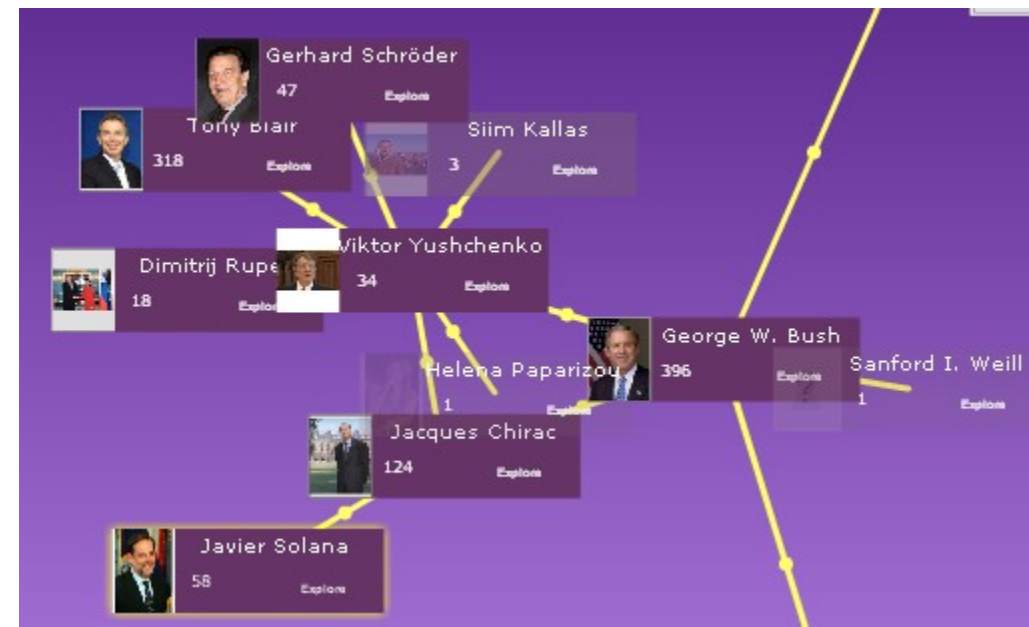
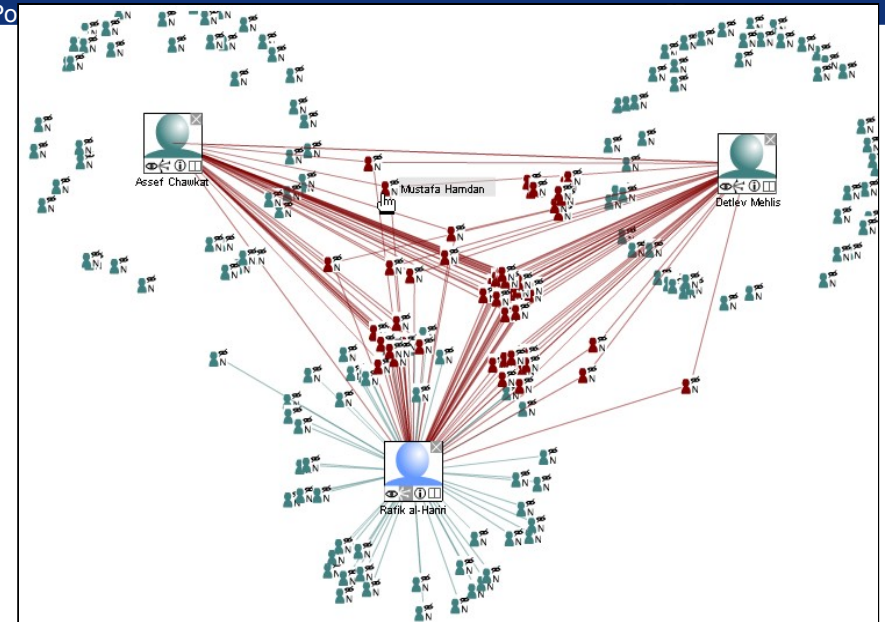
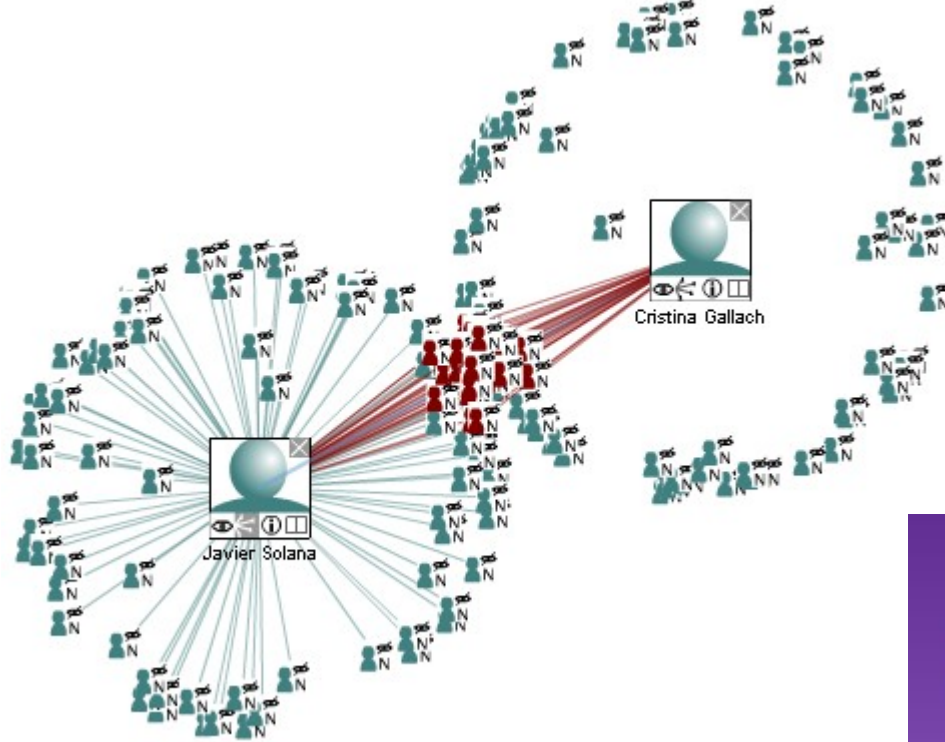
## With weight

	Cristina Gallach		17	17	337	.5993	(de..it)	(24-JUL-04..17-FEB-05)
	Pierre de Boissieu		8	10	93	.4337	(de..fr)	(11-JUN-04..20-AUG-04)
	Jan Kubis		8	10	93	.4337	(de..nl)	(26-NOV-04..11-JAN-05)
	Imad Fallouji		3	3	39	.3398	(de..fr)	(22-JUL-04..23-JUL-04)
	Anatoliy Bliznyuk		3	3	54	.3176	en	(30-NOV-04..02-DEC-04)
	Andriy Mahera		3	3	57	.3142	fr	(04-DEC-04..06-DEC-04)
	Viktor Anpilov		3	3	67	.3045	(en..fr)	(05-DEC-04..06-DEC-04)

Solana's assistant

Solana's spokesperson

- Current visualization:



- **Syntax-based relationship**
- **Quotation-based relationship**

## Syntax-based relationship (in political analyst tasks)

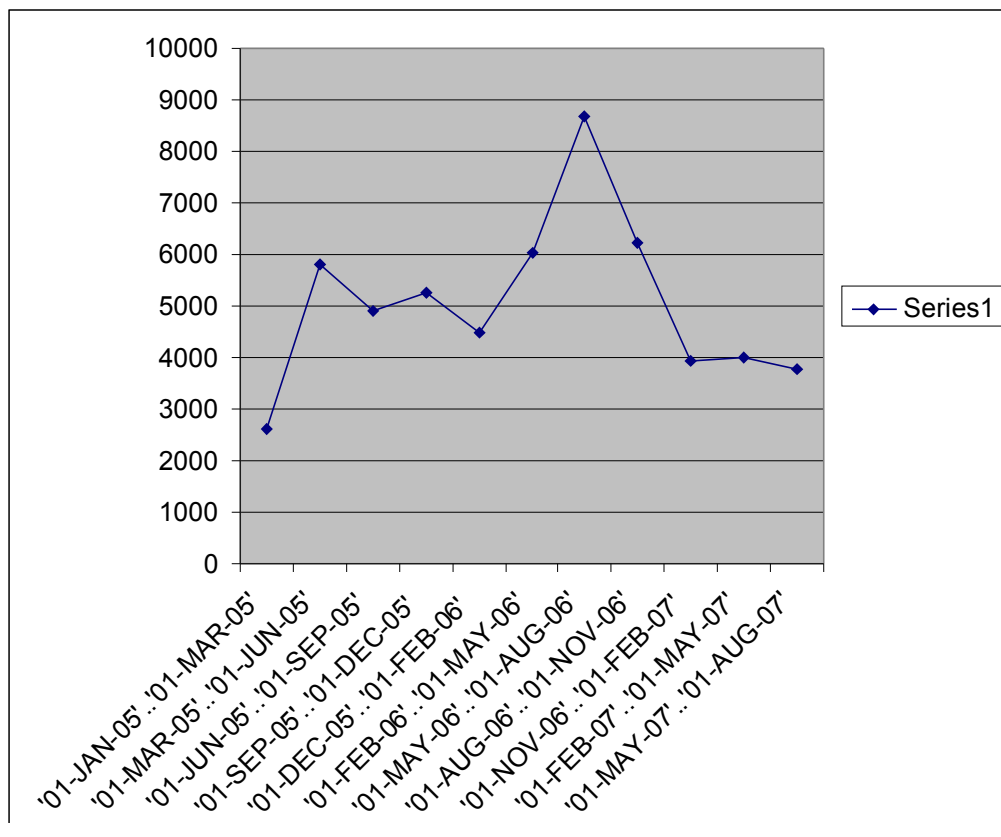
- For the relation “PERSON1 meets PERSON2” we use the syntactic patterns
  - $X \leftarrow \text{subj}-(\text{to meet})-\text{obj} \rightarrow Y$  (Meeting)
  - $X \leftarrow \text{subj}-(\text{married})-\text{obj} \rightarrow Y$  (Family)
- **Then we learn new patterns**
  - **We parse articles**
- **We extract qualified relationships**
  - **Social network, directed graph**

- From a “seed” pattern, e.g.  $X \leftarrow \text{subj}-(\text{to meet})-\text{obj} \rightarrow Y$
- Match these templates against the news clusters in the corpus. Each pair of person names which fill the slots  $X$  and  $Y$  is called an *anchor pair*.
- From “*Bush met the Prime Minister Hamid Karzai*”, the algorithm will extract the anchor pair ( $X:\text{Bush}; Y:\text{Hamid Karzai}$ )
- For each extracted anchor pair, search in the same cluster all the sentences where both names of the anchor pair occur
- From all the sentences in which at least one anchor pair appears, learn syntactic pattern using our pattern-learning algorithm similar to the General Structure Learning algorithm (GSL) described in (Szpektor et.al. 2006)
- Example:  $X \leftarrow \text{subj}-(\text{have dinner})-\text{with} \rightarrow Y$
- Each pattern obtains as a score the number of different anchor pairs which support it

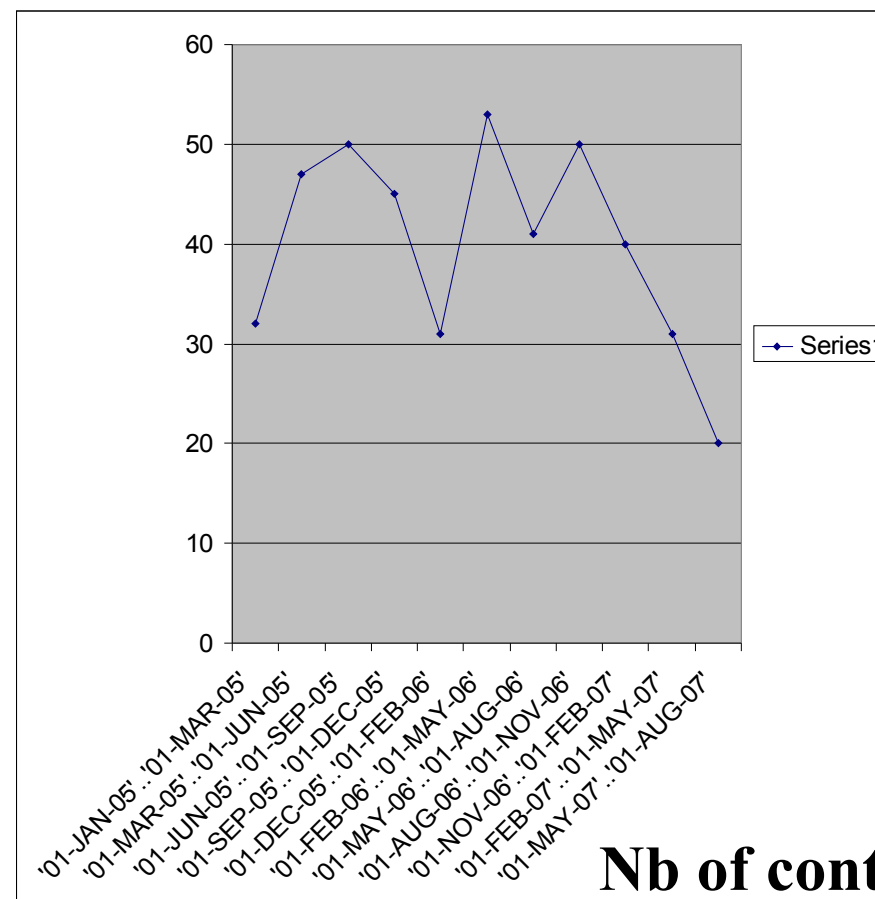
- We extracted Millions of relations for a two-year period of English News
- Comparing Eigen-vector centrality measure (used in PageRank) to estimate the importance of a person
- In one month period oct 2006
- Condoleezza Rice ranked higher because she was on an important tour in middle east

<b>Rank</b>	<b>Eigen vector centrality ranking</b>	<b>Frequency based ranking</b>
1	Condoleezza Rice	George Bush
2	George W. Bush	Tony Blair
3	Vladimir Putin	Condoleezza Rice
4	Ehud Olmert	Nouri al-Maliki
5	Tony Blair	Saddam Hussein

- Use the number of meetings/contacts:
  - Observations of leaders, time period: 01/01/2005-01/08/2007
  - The number of contacts is a good indicator
  - In this experiment it can sometimes be considered more reliable than frequency



**Ex: Condoleezza Rice: Frequency**



**Nb of contacts**



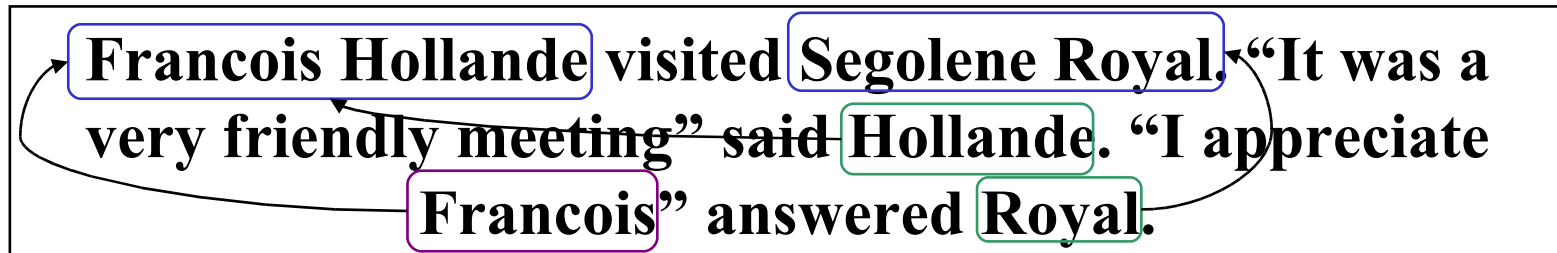
We extract quotations from news articles (in 16 languages)

- Quotation markers
- Reporting verbs
- Modifiers
- Determiners
- Trigger for person
- (from)Person name
- (about) entity

«**L**'esprit de revanche est parfait avant une finale si le **Milan AC**  
parvient à canaliser ses émotions ... **»**, **a dit hier le Néerlandais Gullit**

<http://langtech.jrc.it/entities/quotes/>

- When a name appears in an article, we also lookup parts-of-name

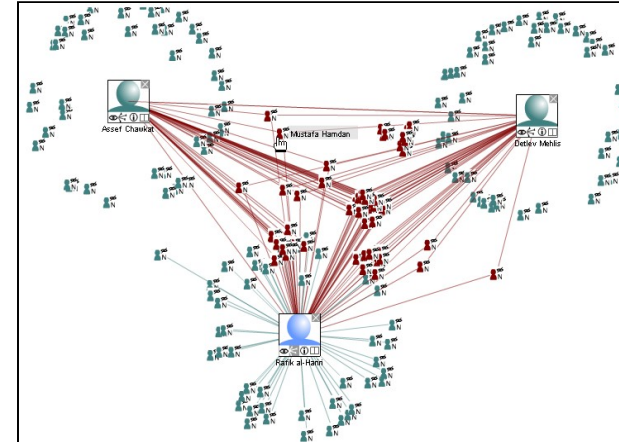


- Once a quote has been identified we look if it does not contain a known “part-of-name”, if so we also record in the database the “mentioned” person
- A person mentioning another person in his/her reported speech creates a link
- We export all “quoted links” from the past month and build a social network



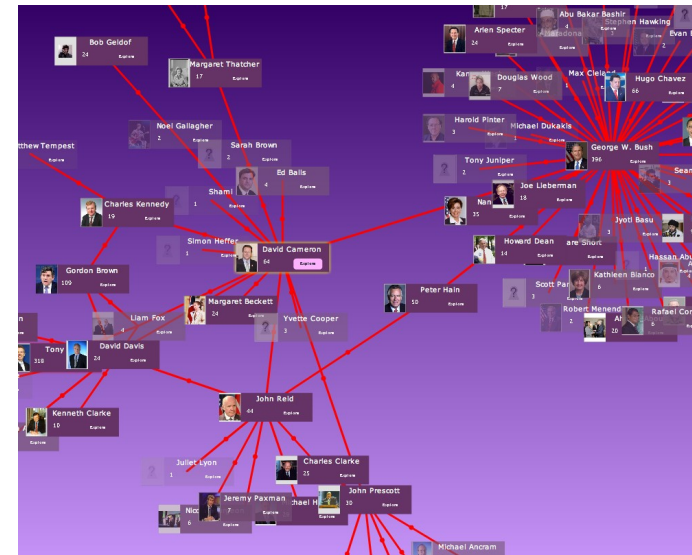


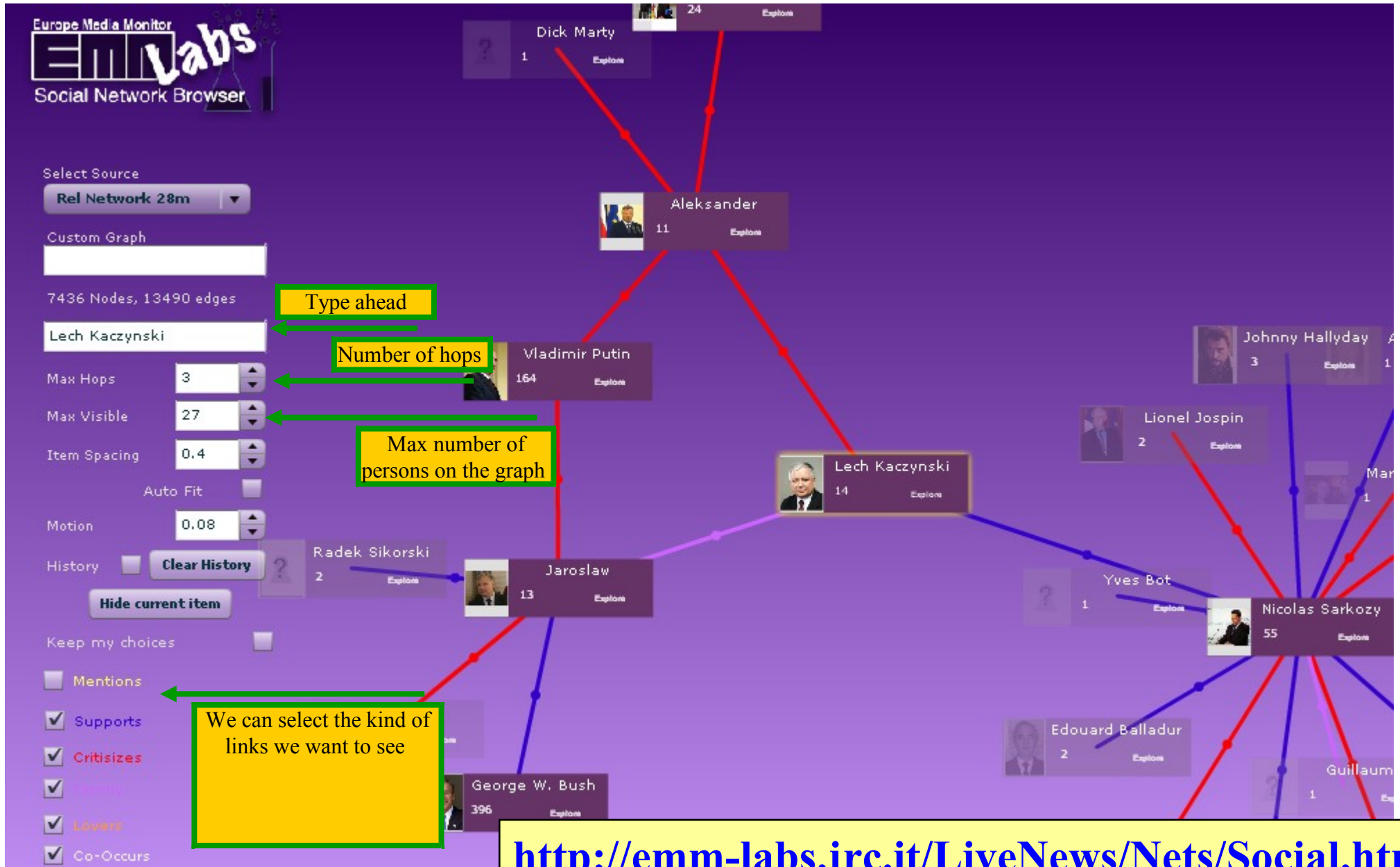
- Flash interactive maps
  - Our first approach
  - Online since 4 years
  - Cannot really visualize many persons



- Experimental system: Spring Graph

- Rich Internet Application client
- Based on Adobe Flex





The screenshot displays the 'Europe Media Monitor Labs Social Network Browser' interface. The main area shows a network graph with nodes representing individuals and edges representing relationships. Nodes include Dick Marty (1), Aleksander (11), Vladimir Putin (164), Lech Kaczynski (14), Jaroslaw (13), Radek Sikorski (2), George W. Bush (396), Johnny Hallyday (3), Lionel Jospin (2), Yves Bot (1), Nicolas Sarkozy (55), Edouard Balladur (2), and Guillaume (1). The interface includes a search bar with 'Lech Kaczynski' entered, and various filters: 'Max Hops' (3), 'Max Visible' (27), 'Item Spacing' (0.4), 'Motion' (0.08), and a 'History' section with 'Clear History' and 'Hide current item' buttons. A list of link types is shown on the left: Mentions (unchecked), Supports (checked), Criticizes (checked), Loves (checked), and Co-Occurs (checked). Annotations with green boxes and arrows point to the search bar ('Type ahead'), the 'Max Hops' control ('Number of hops'), the 'Max Visible' control ('Max number of persons on the graph'), and the link type list ('We can select the kind of links we want to see').

Europe Media Monitor Labs  
Social Network Browser

Select Source  
Rel Network 28m

Custom Graph

7436 Nodes, 13490 edges

Type ahead

Number of hops

Max number of persons on the graph

We can select the kind of links we want to see

Lech Kaczynski

Max Hops 3

Max Visible 27

Item Spacing 0.4

Auto Fit

Motion 0.08

History Clear History

Hide current item

Keep my choices

Mentions

Supports

Criticizes

Loves

Co-Occurs

Dick Marty 1

Aleksander 11

Vladimir Putin 164

Lech Kaczynski 14

Jaroslaw 13

Radek Sikorski 2

George W. Bush 396

Johnny Hallyday 3

Lionel Jospin 2

Yves Bot 1

Nicolas Sarkozy 55

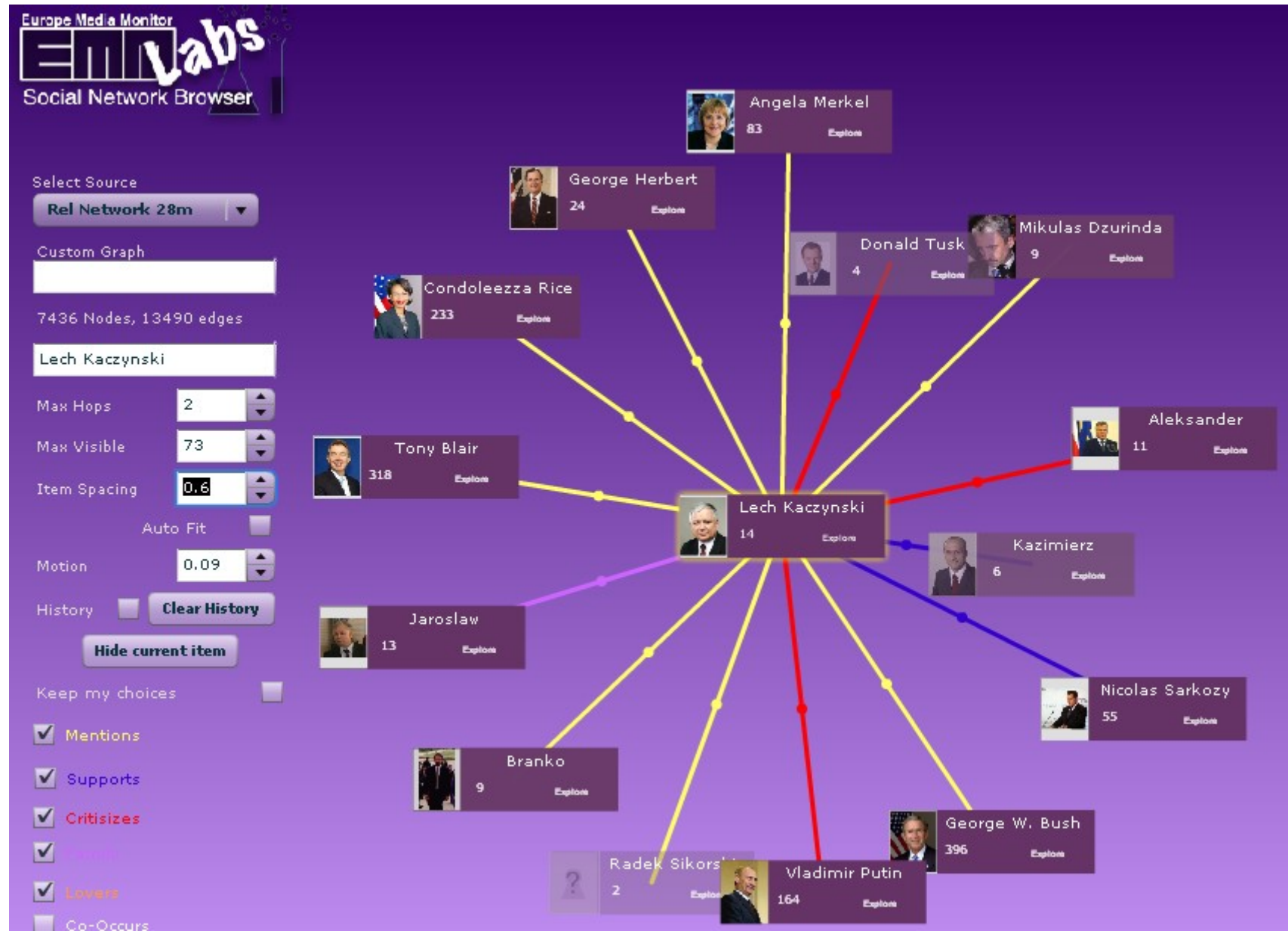
Edouard Balladur 2

Guillaume 1

<http://emm-labs.jrc.it/LiveNews/Nets/Social.html>



## Only direct links (hops = 2)



- We have enormous data (texts)
- Generalization:
  - Meta-data can be derived from texts (ex: entities)
  - Meta-information can be derived from meta-data (ex: relationship)
  - Compact information can be derived from meta-information (ex: social network)
  - And more... (ex: use social networks to study trends...)
- Various approaches to build Social networks
- How to exploit them? What for?
  - There is a need: Currently 2000 hits/day on our “old” social network browser
  - Advertising the technology: Visualization
  - We do not provide ready made socio-political analysis
  - But we offer a nice tool for analysts



Social Network browser <http://emm-labs.jrc.it>

NewsBrief <http://press.jrc.it/>

NewsExplorer <http://press.jrc.it/NewExplorer/>

<http://langtech.jrc.it/picNews.html>

- Contacts: [Bruno.Pouliquen@jrc.it](mailto:Bruno.Pouliquen@jrc.it)
- [Hristo.Tanev@jrc.it](mailto:Hristo.Tanev@jrc.it) (syntax-based patterns)
- [Martin.Atkinson@jrc.it](mailto:Martin.Atkinson@jrc.it) (visualization)