

Journées francophones sur les Entrepôts de Données et l'Analyse en ligne  
**EDA 2015**  
Bruxelles, 2-3 juin 2015

# On-Line Analytical Processing on Graphs generated from Social Network data

**Lilia Hannachi**, Omar Boussaid, **Nadjia Benblidia**, Fadila Bentayeb

LRDSI Laboratory, University of Blida, Algeria  
ERIC Laboratory, University of Lyon 2, France



- Motivation**
- Social Graph Cube**
- Social Graph Cube Lattice**
- Experimental Study**
- Conclusion**

# Motivation

---

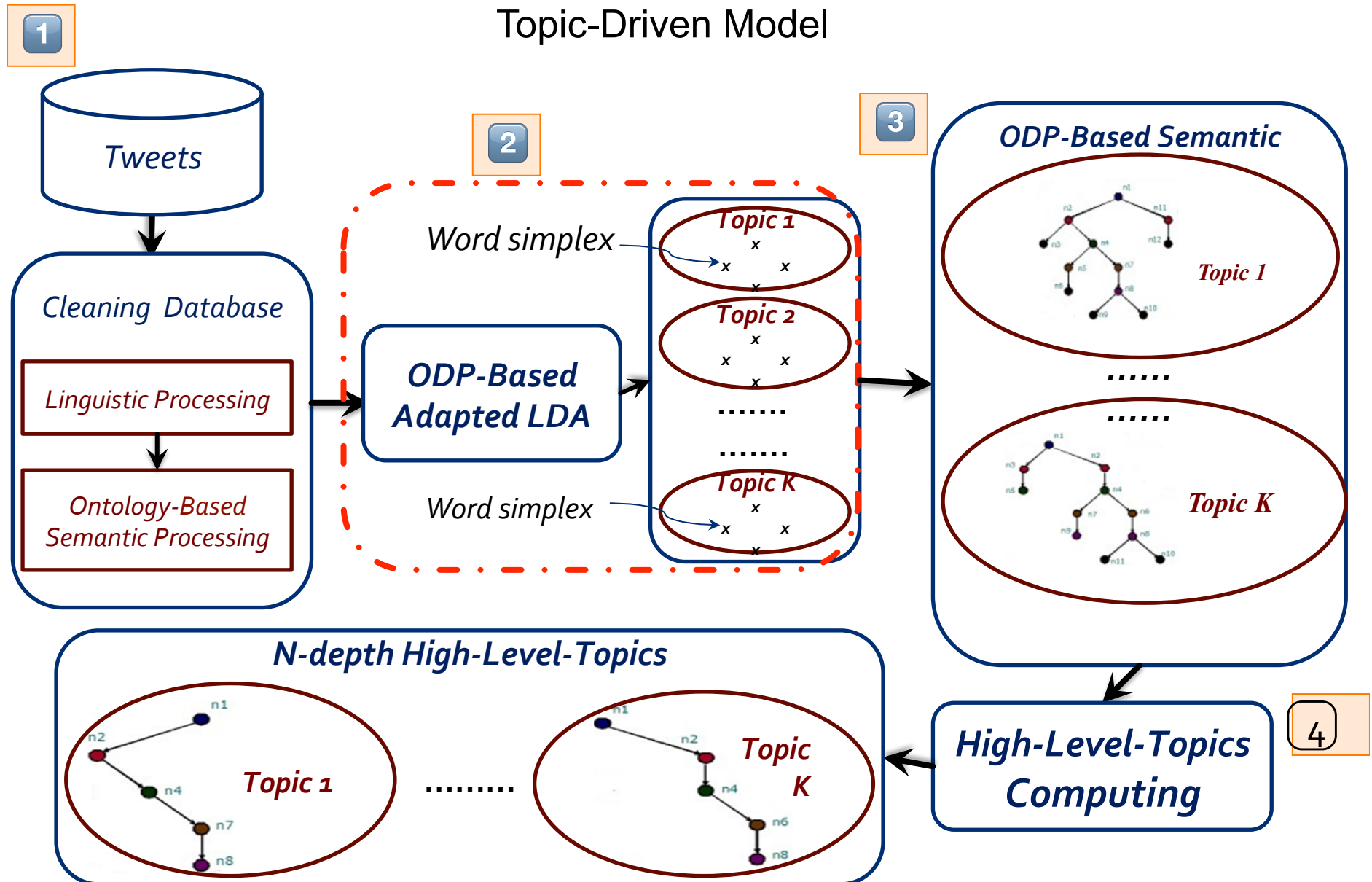
- ❑ Over the last few years, social network sites have been quickly increasing and most particularly within the last ten years.
- ❑ The participation of people in these sites plays a crucial role by publishing real time information about their personal views and interests.
  - How identify the most interesting information among this massive amount of data that is continuously being generated over a period of time.
  - Social networks abundantly feed the big data

# Motivation

---

- ❑ On-Line Analytical Processing represents a powerful and flexible tool to mine and analyze data deeply.
  - It offers analysts the ability to analyze quickly and navigate through the data from different perspectives and with multiple granularities.
  
- ❑ It is more and more important to analyze the social data by using OLAP technique
  - OLAP technology does not consider the different kinds of relationships among individual data tuples.
  - OLAP tools face great challenges for analyzing unstructured data such as ***the social user-generated content.***

# Previous work

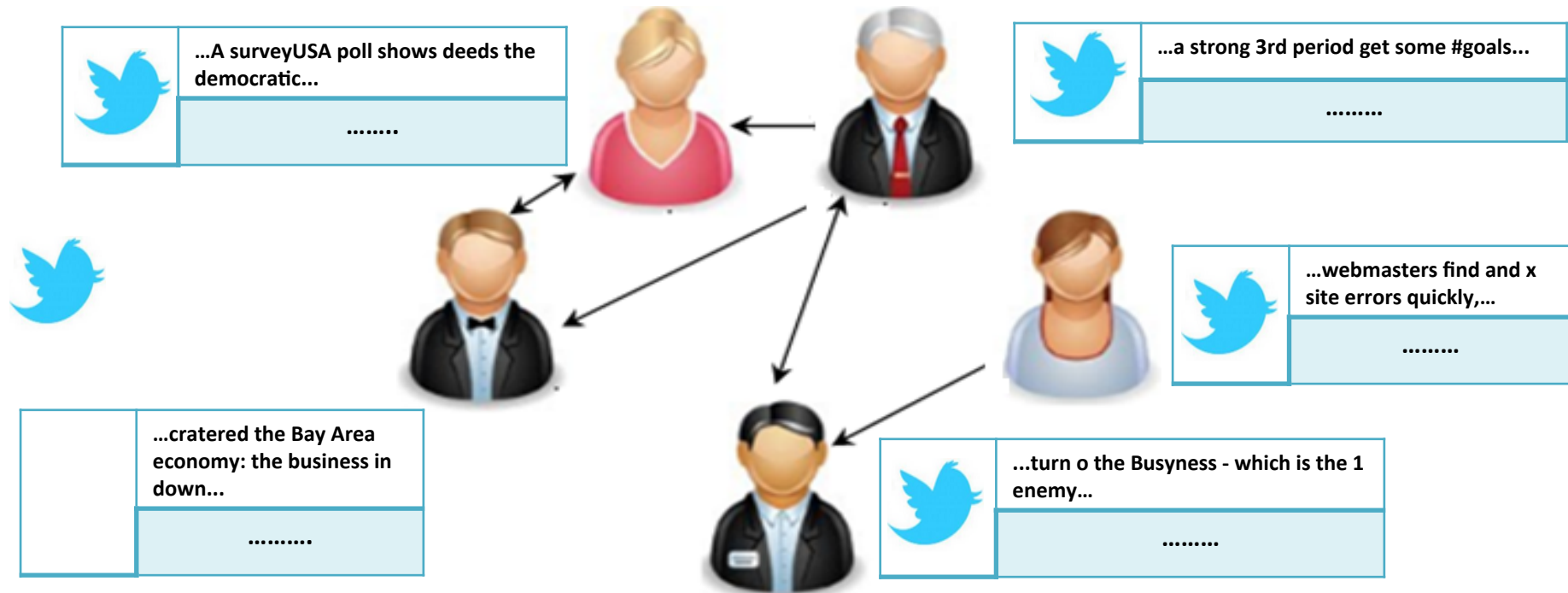


# Social Graph Cube

---

- ❑ ***Social Graph Cube*** is a data warehousing model that can explore and extract the pertinent knowledge hidden in the social network services.
- ❑ ***Social Graph Cube*** permits the decision makers to interactively analyze and manage **structured data, topological structure** and **unstructured user-generated content** according to several perspectives and with different granularities.
- ❑ ***Social Graph Cube*** represents data as heterogeneous information graphs to capture much richer and comprehensive illustration than traditional OLAP cube.

# Social Graph Cube



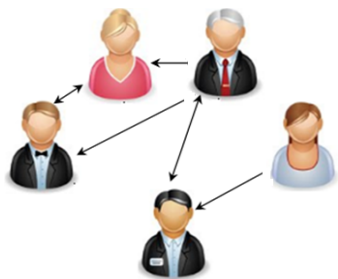
## Associated information with the user generated content:

- User identifier.
- Time.
- Location (Longitude, Latitude)
- etc...

# Social Graph Cube

Users	Example of tweets content
U1	... Nothing affects the modern economy and society more than...
U2	... Looking for investors to fund new project – breathalyzer kiosk that allows...
U3	...J ust heard some devastating news... my prayers goes up...
U4	...We are expecting new AMD news early new year also puno resolution we...
U5	... Ne Job Listening: Rep-Retail Sales (Panama City) at Verizon Wireless (Panama City, FL): Responsibilities Yo...
U6	... Huge loss in home values cratered the Bay Area economy: "There's no doubt ... the business is down." Four years a ...
U7	...I nnovation meeting opportunity at an avenue called not-enough-cash...
U8	... Will AI really change our relationship with tech? ... how would it affect interaction design?...
U9	... Virtual Patients Helping Train Student Nurses At Birmingham City University....

ID	TIME	LOCATION	WORD
1	06/06/14	NY	modern
1	06/06/14	NY	economy
1	06/06/14	NY	society
2	10/06/14	MT	look
2	10/06/14	MT	project
3	20/05/14	FL	hear
4	21/05/14	VA	expect
4	21/05/14	VA	AMD
5	15/06/14	CA	job
6	18/06/14	FL	huge
7	23/05/14	CA	innovation
7	23/05/14	CA	meet
8	05/06/14	NY	AI
...	.....	.....	.....
...	.....	.....	.....
...	.....	.....	.....



The geographical specifications are determined by using:

- The two types of attributes (longitude, latitude).
- The information filled by the user in his profile
- The user's time zone

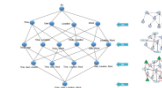
The semantic specifications are determined by using linguistic and semantic knowledge.



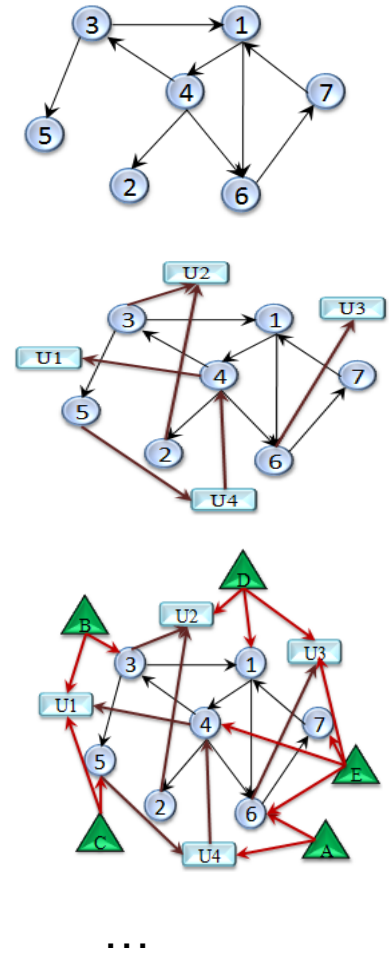
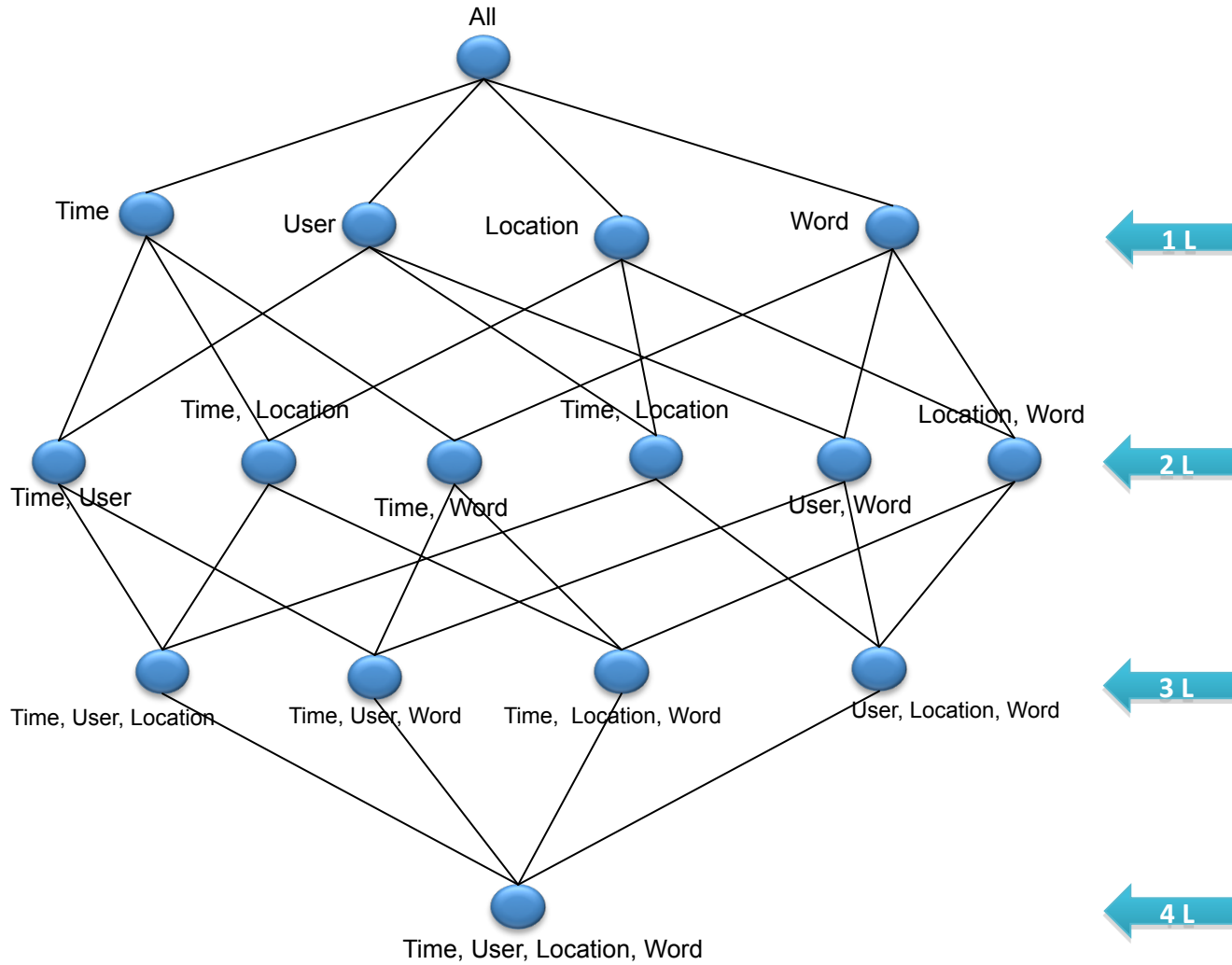
# Social Graph Cube

---

- Given a *multidimensional network*  $\mathbf{N} = (\mathbf{V}, \mathbf{E}, \mathbf{S}, \mathbf{U})$ , the **Social Graph Cube** is obtained by *reorganization of this multidimensional network* in all possible cuboids produced by using the structured data  $\mathbf{S}$  and the user-generated content  $\mathbf{U}$ .
  
- For each cuboid  $\mathbf{C}'$  obtained by  $\mathbf{S}$  and  $\mathbf{U}$ , the measure could be a homogeneous or heterogeneous weighted graph  $\mathbf{G}' = (\mathbf{V}', \mathbf{E}', \mathbf{W}_{V'}, \mathbf{W}_{E'})$  w.r.t.  $\mathbf{C}'$ .
  - In the  $\mathbf{G}'$ , the  $\mathbf{V}'$  is either a simple set of vertices or a set of condensed vertices. The  $\mathbf{E}'$  represents the set of edges illustrated in the graph  $\mathbf{G}'$ .
  - $\mathbf{W}_{V'}$ ,  $\mathbf{W}_{E'}$  are the list of weights associated with each vertex and each edge in the weighted graph, respectively.



# Social Graph Cube Lattice



# The First Level in the Lattice

---

- The cuboids generated in this level are represented with *homogeneous weighted graphs*, where only *1 dimension* is selected.
  
- It can answer some complex queries that could be asked on a multidimensional network such as :
  - *What is the semantic network structure between the several users ?*
  - *What is the semantic relationship between the most mentioned words in the social user generated content ?*

# Social Graph Cube

---


## Construction approach of a Social Graph Cube

- ❑ *Step 1:* to aggregate all the social user generated content sent by the same user, transmitted from the same location or produced at the same time interval.
- ❑ *Step 2:* to calculate the *semantic distance* between each social entity (users, or location, or time intervals)
- ❑ *Step 3:* to compute the weight associated with each social entity (i.e. *vertex*) in the weighted graph.
- ❑ *Step 4:* to construct the semantic weighted graphs associated with each cuboid in the first level of *social graph cube* lattice.

# The First Level in the Lattice

---

- The construction of the multidimensional graph in the case of **word dimension** is achieved by using the following process:

- Distance between  $u_i$  and  $u_j$  is a symmetric **Kullback-Leibler Divergence** 
- Determine the closeness between words by using the well-founded semantic measure *Normalized Google Distance*:

$$NGD(w_i, w_k) = \frac{\max\{\log f(w_i), \log f(w_k)\} - \log f(w_i, w_k)}{\log P - \min\{\log f(w_i), \log f(w_k)\}}$$

- Calculate the weight of words by adapting the closeness centrality measure.

$$Centrality(word_i) = \frac{1}{\sum_{i \neq j} distance(i, j)} + Mean(TDIDF(word_i))$$

- Evaluate the existence of edges between words according to the distance values.

# The First Level in the Lattice

---

- Distance between  $u_i$  and  $u_j$  is a symmetric ***Kullback-Leibler Divergence***

$$dist(i, j) = \sum_{t=1}^T [\theta_{it} \log \frac{\theta_{it}}{\theta_{jt}} + \theta_{jt} \log \frac{\theta_{jt}}{\theta_{it}}]$$

- Adapted to words

$$dist_{s1}(user_s, user_l) = \sum_{k=0}^K \left( TFIDF_s(w_k) \log \frac{TFIDF_s(w_k)}{TFIDF_l(w_k)} + TFIDF_l(w_k) \log \frac{TFIDF_l(w_k)}{TFIDF_s(w_k)} \right)$$

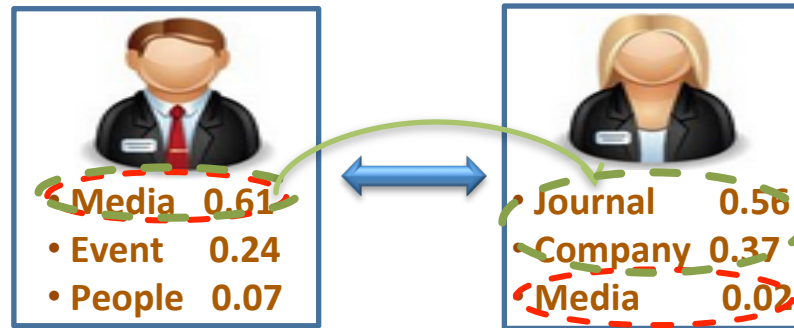


# The First Level in the Lattice

---

- The construction of the multidimensional graph in the case of *user*, *time* or *location* dimensions is achieved by using the following process:
  - Aggregate and clean all the social user-generated content sent by the same user, transmitted from the same location or produced at the same time interval
  - Calculate the semantic distance between the several entities as the *Kullback-Leibler* divergence between the *TF-IDF* of the *top-K* most representative words.

# The First Level in the Lattice



$$dist_{s1}(user_s, user_l) = \sum_{k=0}^K \left( TFIDF_s(w_k) \log \frac{TFIDF_s(w_k)}{TFIDF_l(w_k)} + TFIDF_l(w_k) \log \frac{TFIDF_l(w_k)}{TFIDF_s(w_k)} \right)$$

Define the semantic relationship between the different words



$$dist_{s2}(user_s, user_{sl}) = dist_{s1}(user_s, user_{sl}) + \frac{1}{2} \left( \sum_{k=0}^K \sum_{i \in \{K-k\}} TFIDF_s(w_k) \log \frac{TFIDF_s(w_k)}{TFIDF_l(w_k)} + TFIDF_l(w_i) \log \frac{TFIDF_l(w_i)}{TFIDF_s(w_k)} + NGD(w_k, w_i) \right)$$



# The First Level in the Lattice

---

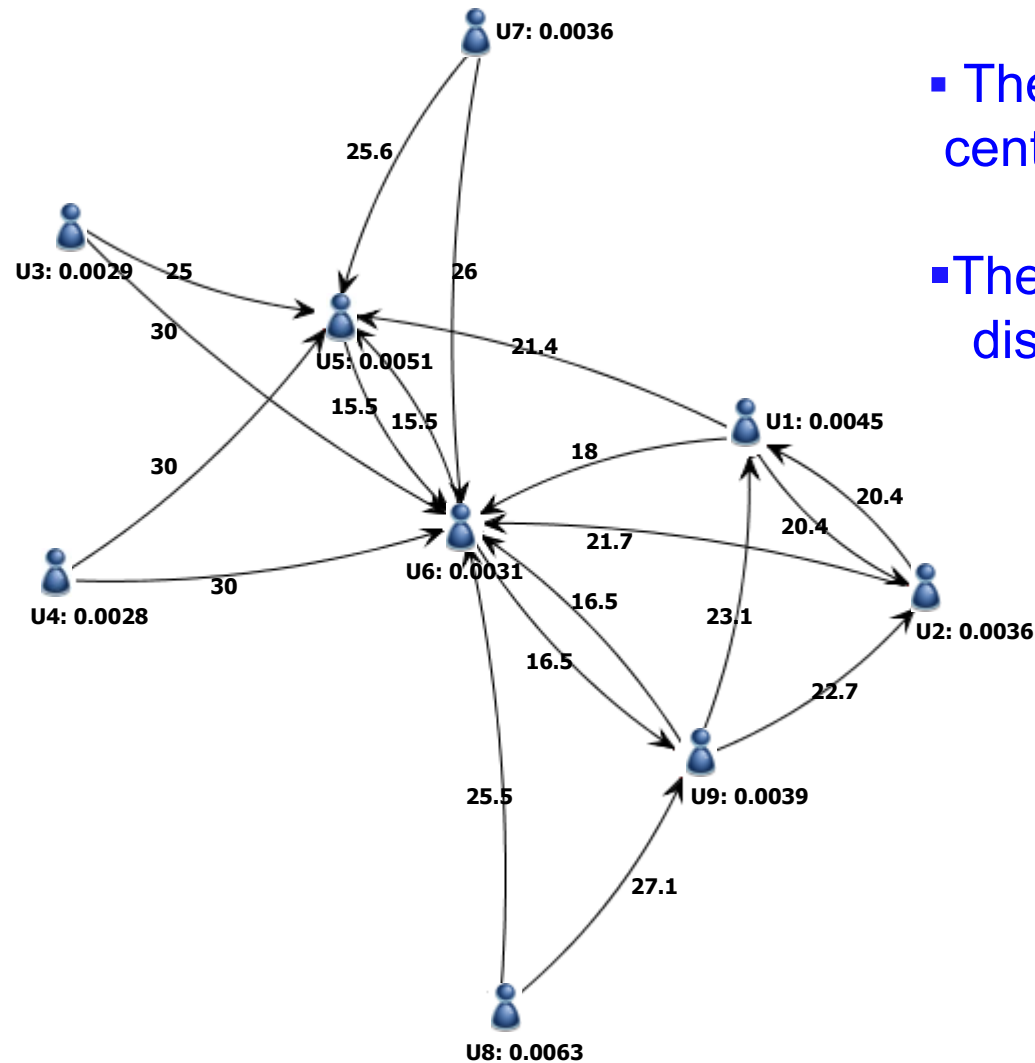
- Compute the **weight** associated with each social entity in the weighted graph.

$$Centrality(s) = \frac{1}{\sum_{s \neq l} distance(s, l)}$$

- Construct the semantic weighted graphs by creating an edge between entities if the semantic distance between them  $\leq$  *threshold* ( $\epsilon$ )
  - The *threshold* ( $\epsilon$ ) can be defined by either the end-users or by the mean associated with the selected entity:

$$mean_s = \frac{\sum_{l=0}^L dist_{s2}(user_s, user_l)}{L}$$

# The First Level in the Lattice



- The vertex weight is the closeness centrality.
- The edge weight is the semantic distance between entities.

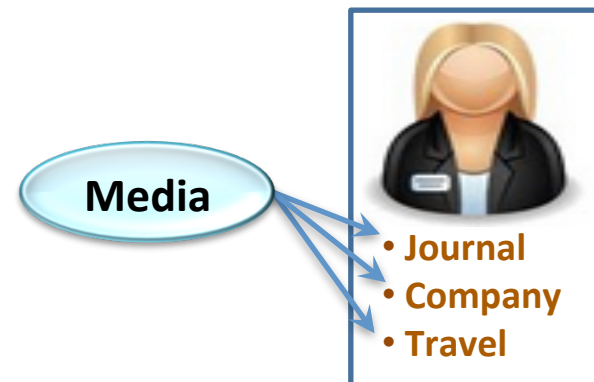
FIG. 1: An instance of the weighted user-user graph



# The Second Level in the Lattice

- The second level in the Social Graph cube lattice represents the first step to generate heterogeneous weighted graphs by accumulating two different multidimensional spaces.
- In the case of heterogeneous graphs produced through the semantic dimension word, the semantic distance is calculated by:

$$\text{dist\_word}(user_s, w) = \frac{\sum_{k=1}^K NGD(w_k, w)}{K}$$



# The Second Level in the Lattice

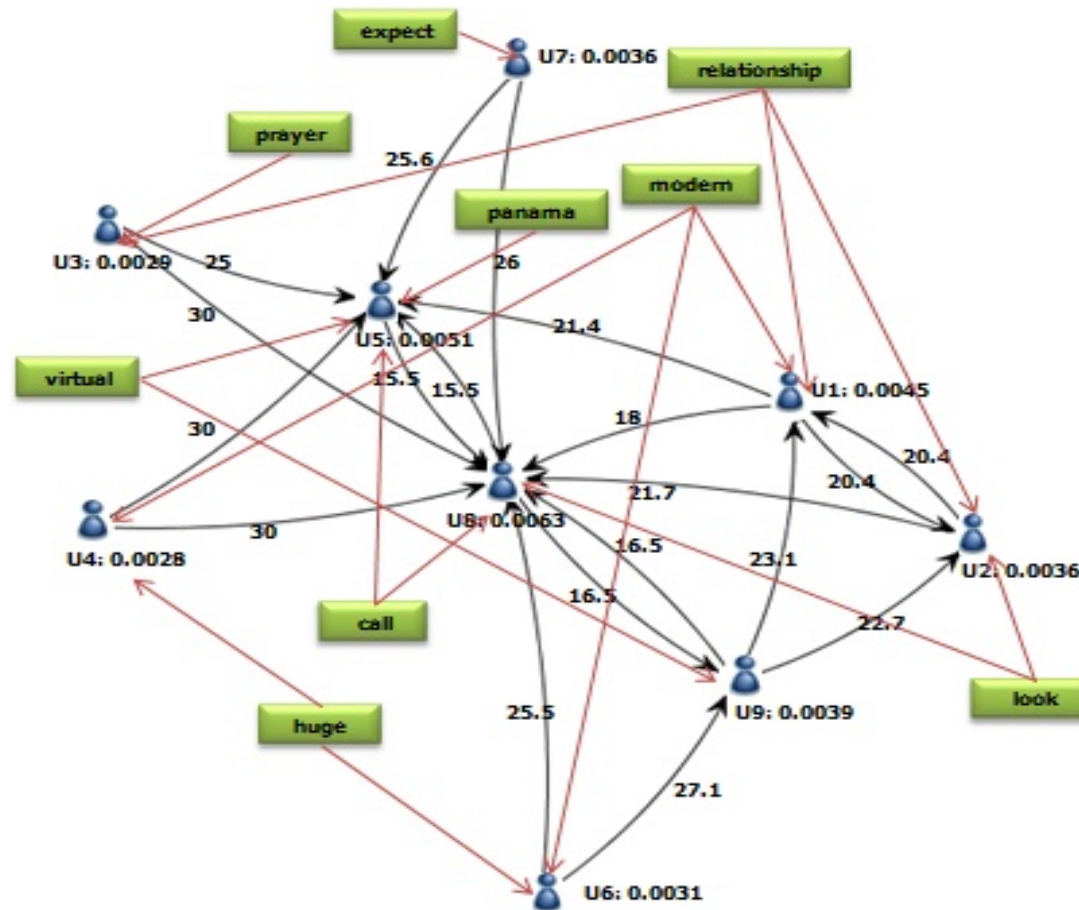
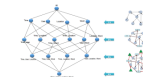


FIG. 2: The heterogeneous weighted user-word graph



# The Third and Fourth Levels

---

- According to the number of entities involved in the end-users requests, heterogeneity of graphs generated in these levels is increased.
- The process utilized in these levels is similar to the process illustrated in the second level. The only difference is the number of the selected multidimensional spaces.
- These heterogeneous weighted graphs leverage the rich semantic knowledge hidden in the massive social structure.

# The Social Graph Cube Aggregations

- The proposed Social Graph cube integrates OLAP technologies, community extraction methodologies and data mining clustering in a unified approach in order to represent the social data in a summarized visualization.
- All possible aggregations are determined by using both structural and topological data associated with the social network.
- The popular modularity measure to define the best result.

$$Q = \frac{1}{m} \sum_{i,j \in V} \left( A_{ij} - \frac{k_i^{out} k_j^{in}}{M} \right) \delta(C_i, C_j)$$

# Experimental Study

---

Protocol :

**Dataset**: a corpus of 4 million tweets ; 3 000 relevant users (# followers, #retweets..)

**Objective**: observe tweets location from semantic, geographic and temporal axes

Communities extraction process:

1°) Compute the topological and the semantic distances using the length of the shortest path and the semantic distance.

2°) Agglomerative strategy is utilized to extract the users clusters or location clusters by using the content and topological distance computed previously.

3°) The extracted clusters are evaluated to get the best result.  
The popular modularity measure that evaluates the extracted communities.

# Experimental Study



<b>C1:</b> United States, China, Japan, South Korea	Intellect, assistance, china, system, digital, student
<b>C2:</b> Algeria, Tunisia, Egypt, Syria, Saudi Arabia, Yemen, Lebanon	Democrat, news, religion, internet, facebook, election
<b>C3:</b> Spain, Italy, Greece	Security, crisis, industry, company, business, woman
<b>C4:</b> Namibia, South Africa, Zimbabwe	Tourism, nature, transport, photo, justice, health

FIG. 3 – Compressed vision with the top used words of semantically related countries



# Conclusion

---

- We proposed a new multidimensional model, *Social Graph Cube*, for efficient and effective exploration of data contained in the multidimensional social network.
  - It breaks the boundaries of the classical OLAP by illustrating the analysis results as homogeneous and heterogeneous weighted graphs.
  - It presents a new method that combines data mining field and OLAP operators to navigate through hierarchies.
- Improve our experimentation and propose new evaluation protocol in the future works.

# OLAP on Graph generated from Social Network data

---

Questions?

