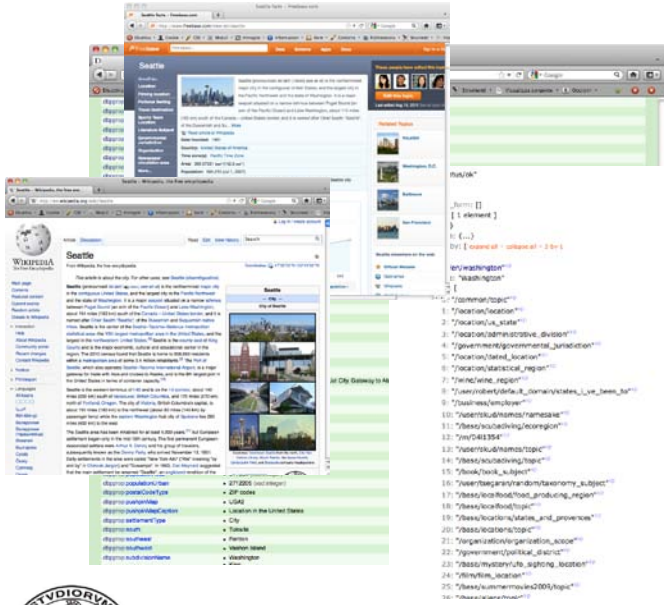
Thematic Exploration of Linked Data

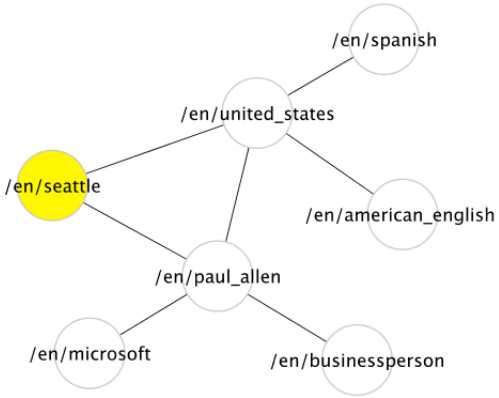
Silvana Castano,
Alfio Ferrara,
Stefano Montanelli
Università degli Studi di Milano
DICo, 20135 Milano, Italy
name.surname@unimi.it

VLDS 2011
@VLDB, Seattle

Linked data are useful…

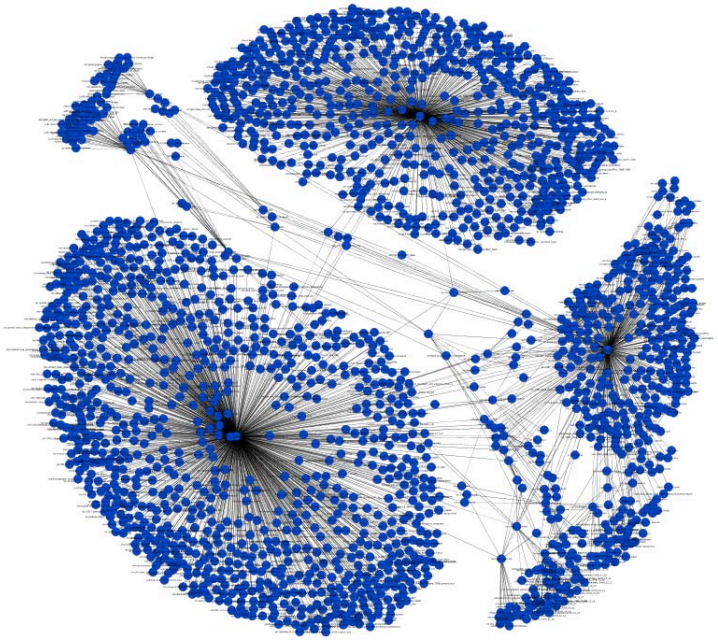
- Linked data are often an invaluable source of information about a target of interest
- They may be used to quickly learn something about the target or even to support the (semi) automatic development of knowledge-intensive (web) applications





…but…

- Learning from linked data usually requires to face a long and loosely-intuitive browsing activity
- It may be hard to discriminate data with respect to their prominence for the target

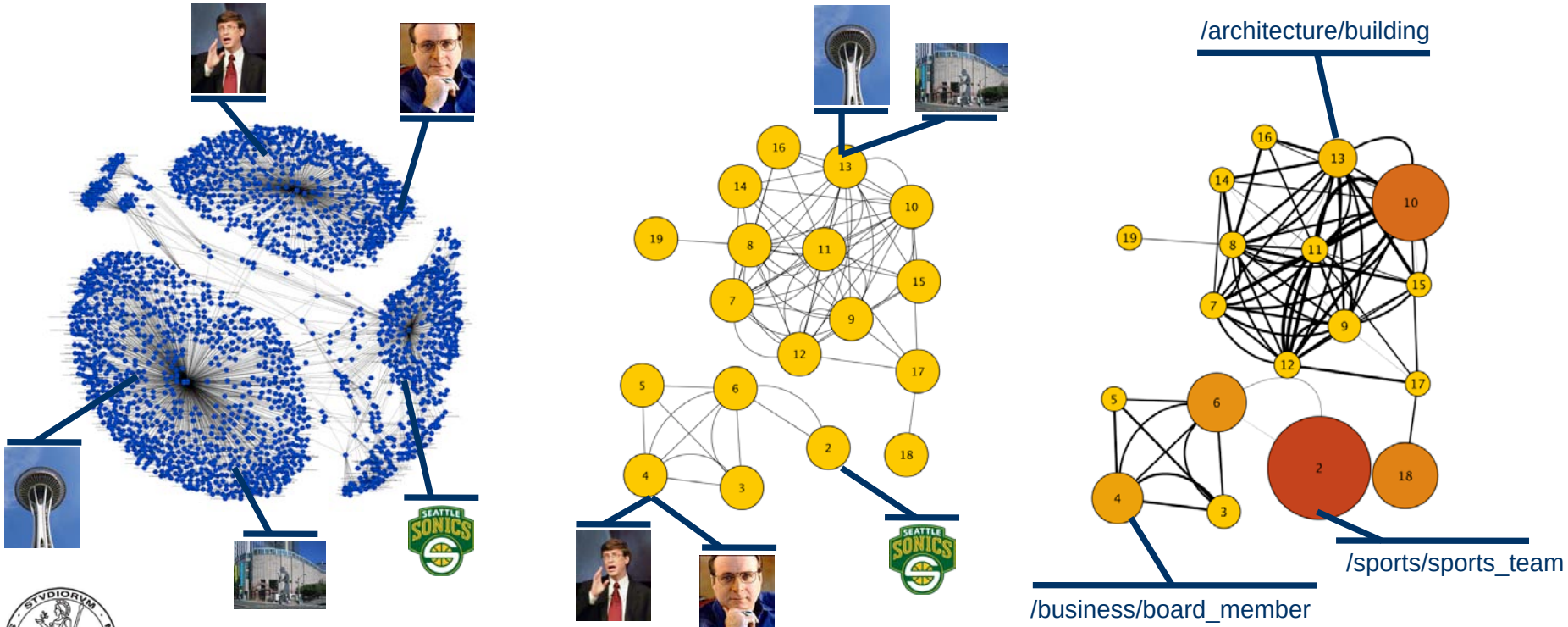


From browsing to thematic exploration

Data acquisition
-Keyword search
-Selection and filtering

Data aggregation
-Matching
-Clustering

Data abstraction
-Essential abstraction
-Prominence & proximity



Data acquisition

- Data acquisition is executed in three steps:
  - Submission to the repository of a search target, i.e. a keyword or a set of keywords

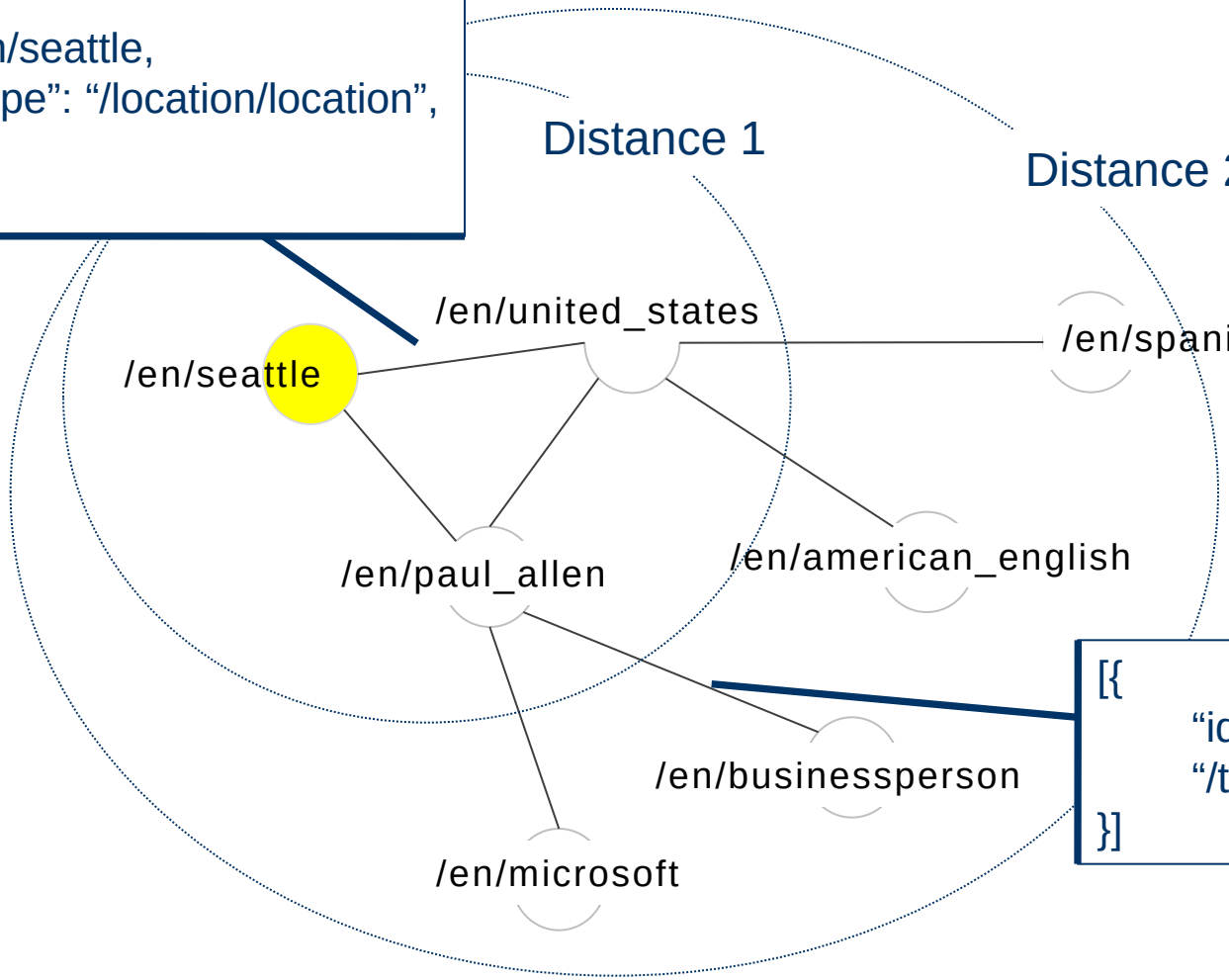
    E.g., we run the keyword “seattle” on Freebase
  - Selection of a seed of interest, i.e. a URI or ID as a starting point for exploration

    E.g., we selected the seed /en/seattle
  - Exploration of the items reachable from the seed within a given distance from the seed

    E.g., we explored the complete set of items directly linked with /en/seattle and some at distance 2

Example of acquisition

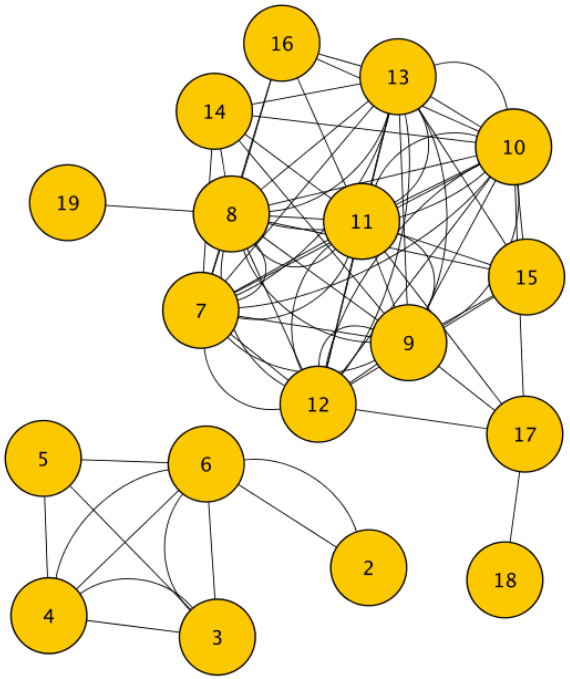
```
[{
    "id": "/en/seattle,
    "/type/type": "/location/location",
    "*": {}
}]
```




```
[{
    "id": "/en/seattle,
    "/type/type": []
}]
```

Data aggregation

- Data aggregation has the goal of grouping together items according to their degree of similarity and mutual linking



- Matching of linked data items
  - Using our matching system HMatch 2.0
- Clustering linked data items
  - Exploiting clique percolation (CPM) [G. Palla et al. *Nature* '05] or other clustering techniques

Matching of linked data

- HMatch 2.0 supports several different matching techniques that can be used in order to evaluate similarity between two linked data items
  - String/term matching (used in the paper)
  - Instance matching, e.g.:

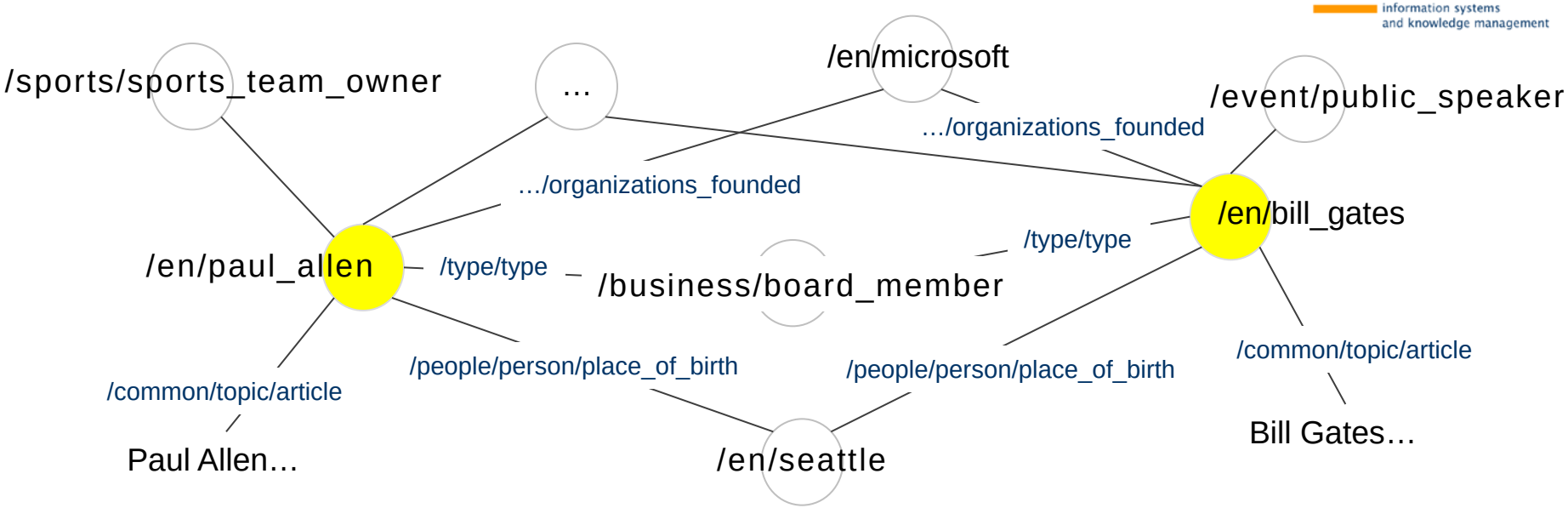
$$sim_{ij} = ws_{ij} + (1 - w)c_{ij}$$

With

$$s_{ij} = \min\left(1, \frac{\left|\{e_l(n_i, n_v) \in G_s : \exists e_m(n_j, n_z), l = m \wedge n_v = n_z\} \cup \{e_l(n_j, n_v) \in G_s : \exists e_m(n_i, n_z), l = m \wedge n_v = n_z\}\right|}{\left|\{e_l(n_i, n_v) \in G_s : \exists e_m(n_j, n_z), l = m\} \cup \{e_l(n_j, n_v) \in G_s : \exists e_m(n_i, n_z), l = m\}\right|}\right)$$

$$c_{ij} = \left(\frac{\left|\{t_i : \exists type(n_i, t_i) \in G_s\} \cap \{t_k : \exists type(n_j, t_k) \in G_s\}\right|}{\left|\{t_i : \exists type(n_i, t_i) \in G_s\} \cup \{t_k : \exists type(n_j, t_k) \in G_s\}\right|}\right)$$

Matching example

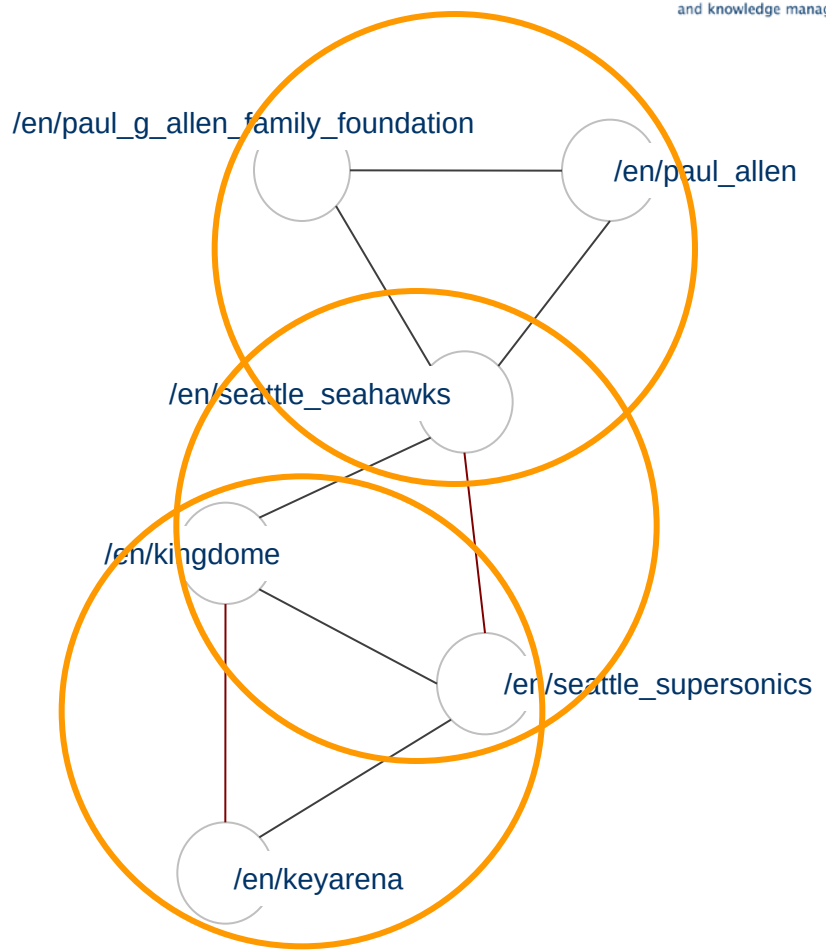


Number of properties in common = 9 (8 with the same value). 4 of 18 types in common.

$$sim = 0.5 \min\left(1, \frac{8}{9}\right) + (1 - 0.5)\frac{4}{18} = 0.62$$

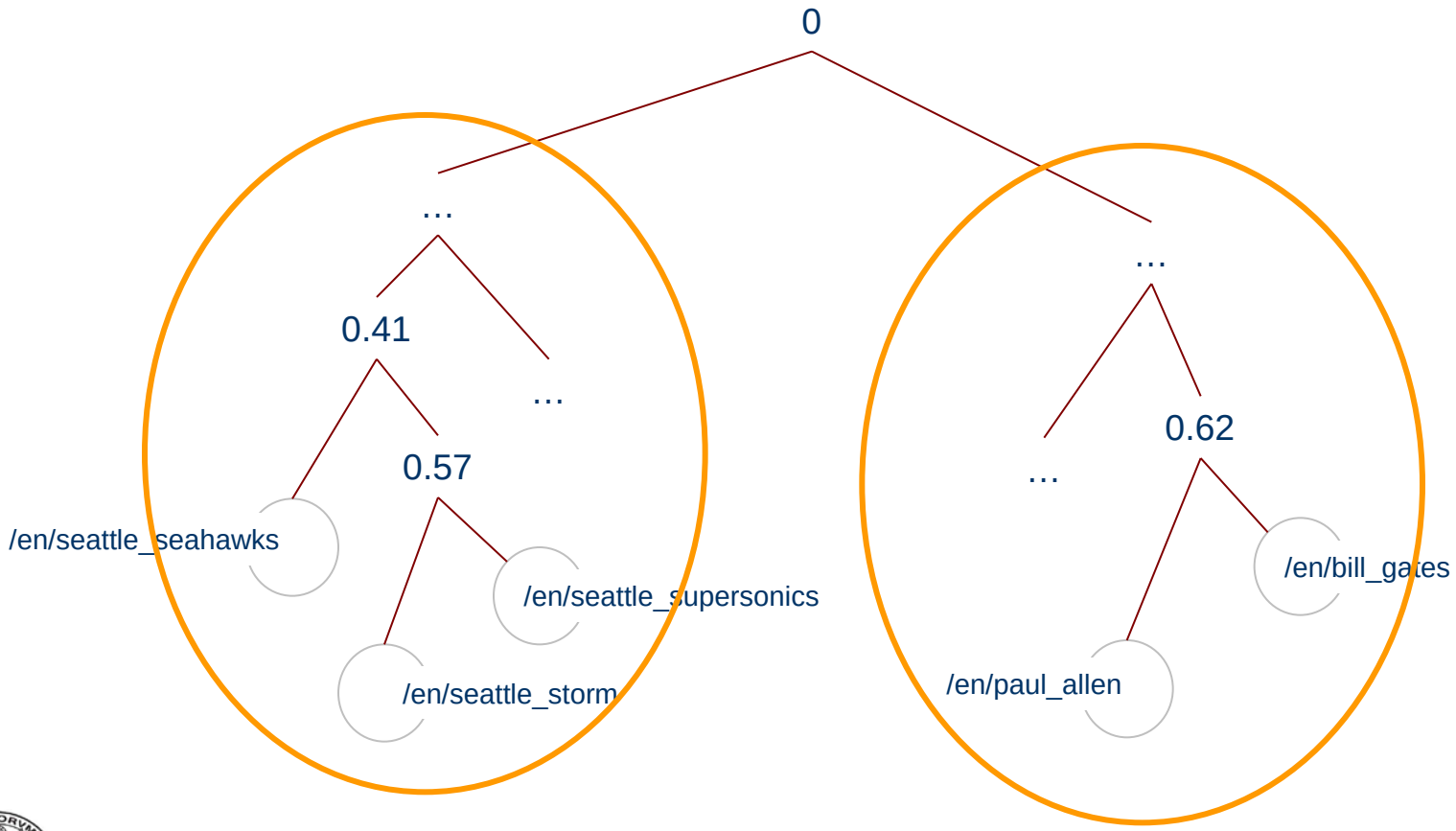
Data clustering

- In order to cluster data, we add mappings produced through matching to the initial resource graph G_s
- We run CPM to find *k*-clique, i.e., complete sub-graph of *k* nodes within the graph G_s+
- Two k-cliques are defined as adjacent if they share k – 1 nodes
- A k-clique-cluster, is defined as the union of all k-cliques that can be reached from each other through a series of adjacent k-cliques



Other approaches to aggregation

- Other clustering methods can be used…



Considerations

- Kind of result:
  - CPM produces a high number of highly overlapping clusters
  - With CPM, clusters can be less homogeneous wrt the topic, but more useful to discover new and interesting item relations
  - HC produces a small number of disjoint clusters
  - With HC clusters are very focused on the topic they represent
- Scalability:
  - CPM is not scalable over 5/6K RDF triples

*in*Clouds: essentials, prominence, and proximity

- In order to build the final *in*Cloud, we transform clusters in essentials
  - An essential represents a collection of linked data items contained in a cluster
  - Each linked data item in a collection is associated with a relevance value
  - The essential itself has a prominence
  - Essentials are labeled with types and keywords
  - Essentials are mutually interconnected by proximity relations

Essential definition

- For each item *j* within an essential e_i, we calculate the relevance r_{ij} of *j* wrt e_i as follows

$$r_{ij} = \frac{\left|\left\{m_{jk} : k \in e_i\right\}\right|}{\left|\left\{m_{jq}\right\}\right|}$$

- Where m_{ab} denotes a mapping between two items a, b with $sim(a,b) > 0$

```
/en/mount_baker                          0.86
/en/madrona_valley_seattle_washington    0.84
/en/bryant_washington                    0.65
/en/haller_lake                          0.65
/en/west_seattle_seattle_washington      0.64
/en/broadmoor_washington                 0.64
/en/ballard                              0.64
/en/south_lake_union                     0.64
/en/northgate_seattle_washington         0.62
/en/university_district                  0.62
/en/hillman_city                         0.62
…
```

Item relevance and prominence

- According to the relevance of each item in an essential, the prominence of an essential e_i is calculated as follows

$$P_i = \frac{2(1 - v_i)d_i}{(1 - v_i) + d_i}$$

- Where d_i is the degree of interconnection among items of e_i (i.e., the density of the graph in e_i)
- v_i is a coefficient of variation, defined as:

$$v_i = \frac{1}{\overline{r_i}} \sqrt{\frac{1}{N_j - 1} \sum_{j=1}^{N_j} \left(r_{ij} - \overline{r_i} \right)^2}$$

Essential descriptors

- Then, we label each essential with a ranked list of types and keywords extracted from types and terminological equipments of linked data items contained in the essential
- The relevance of each type *t* wrt an essential e_i is calculated as follows:

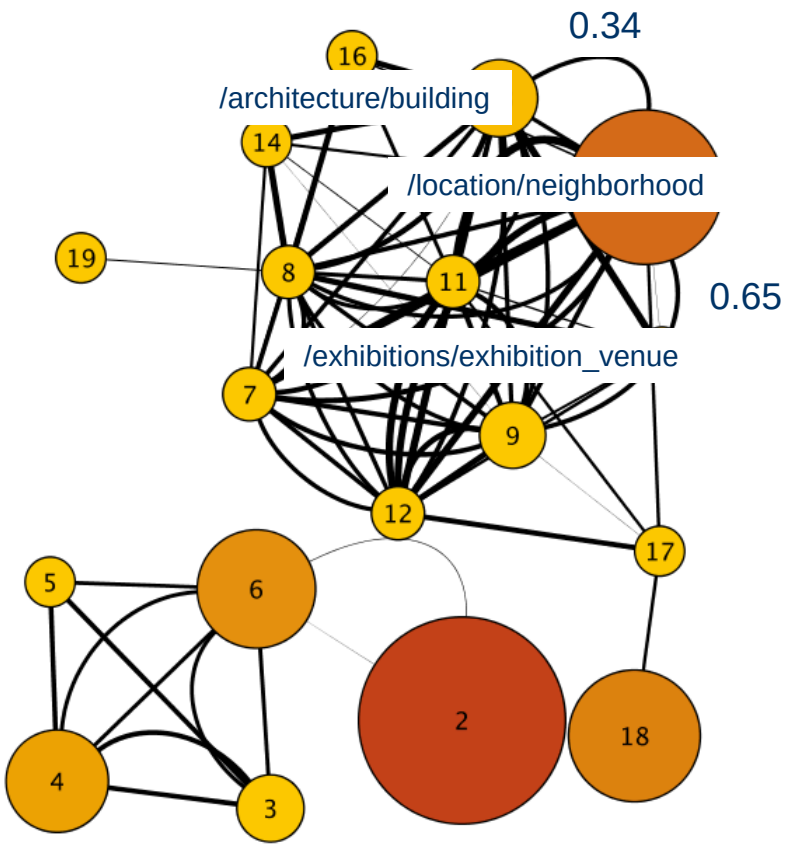
```
/location/neighborhood              0.97
/location/location                  0.084
/geography/geographical_feature     0.06
/geography/body_of_water            0.03
…
```

$$r_{ti} = \frac{\left|\left\{ j \in e_i : \exists type(j,t) \right\}\right|}{\left|\left\{ e_i \right\}\right| \left|\left\{ e_l : \exists type(l,t) \wedge l \in e_l \right\}\right|}$$

Proximity

- The proximity of two essentials e_i and e_j represent their degree of similarity/overlapping
- In particular, proximity is proportional to the number and relevance of similar items in the two essentials as follows

$$X_{ij} = \frac{\sum_{l=1}^{|e_i|} r_{il} : \exists m_{lz}, z \in e_j \sum_{k=1}^{|M_{ij}|} sim^k}{\left(\sum_{l=1}^{|e_i|} r_{il}\right) |M_{ij}|}$$

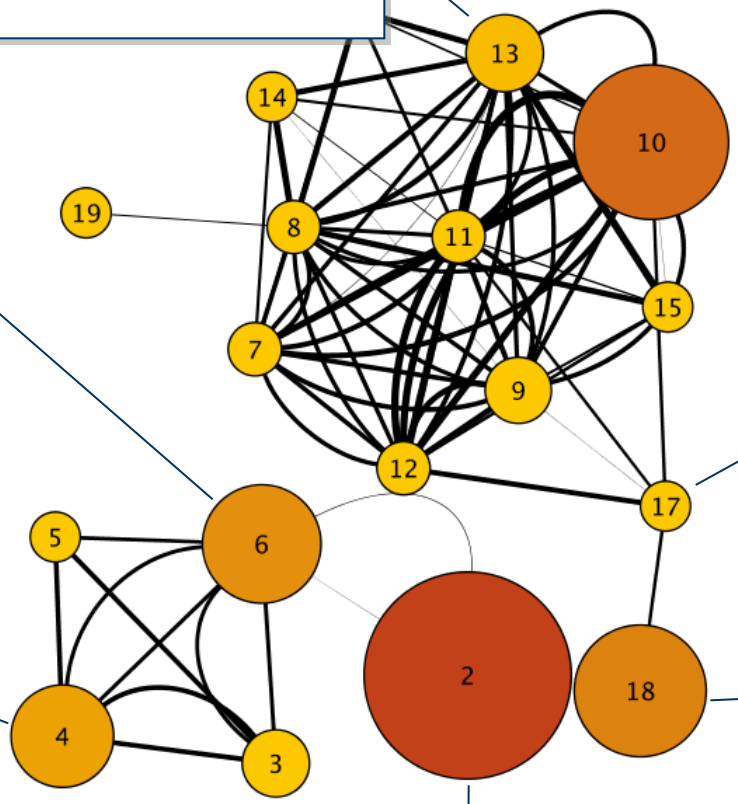


The resulting *in*Cloud

/architecture/building
/en/seattle_art_museum
/en/smith_tower
/en/museum_of_flight
…

/location/neighborhood
/en/mount_baker
/en/madrona_valley_
_seattle_washington
/en/bryant_washington
…

/award/award_nominee
/en/ann_pennington_1950
/en/janice_pennington
/en/erika_christensen
…



/travel/travel_destination
/en/seattle

/business/board_member
/en/bill_gates
/en/doug_isaacson
/en/paul_allen
…

/location/country
/en/united_states
/en/germany

/sports/sports_team
/en/seattle_mariners
/en/seattle_supersonics
/en/seattle_seahawks
…

Preliminary results

- We run a user evaluation of out approach involving a group of students of the Databases course of the Master Degree in Computer Science held at the University of Milan
- Such students had a similar background on Linked Data and Semantic Web, mainly based on a few classes delivered on these topics in the course
- We presented three test cases and asked the students to learn about the three topics with conventional web tools and with *in*Clouds

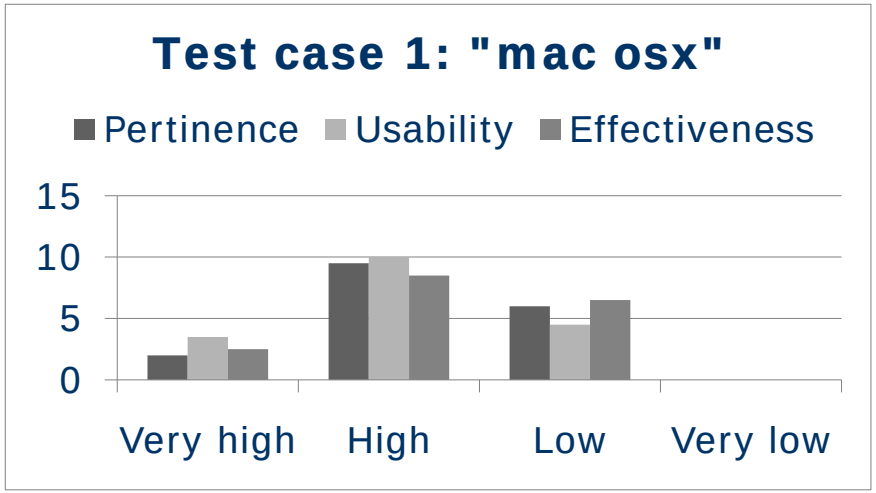
Test case	Target	# items	Data sources
1	“mac osx”	148	Freebase
2	“star wars”	423	Freebase + DBPedia
3	“london olympics”	190	DBPedia

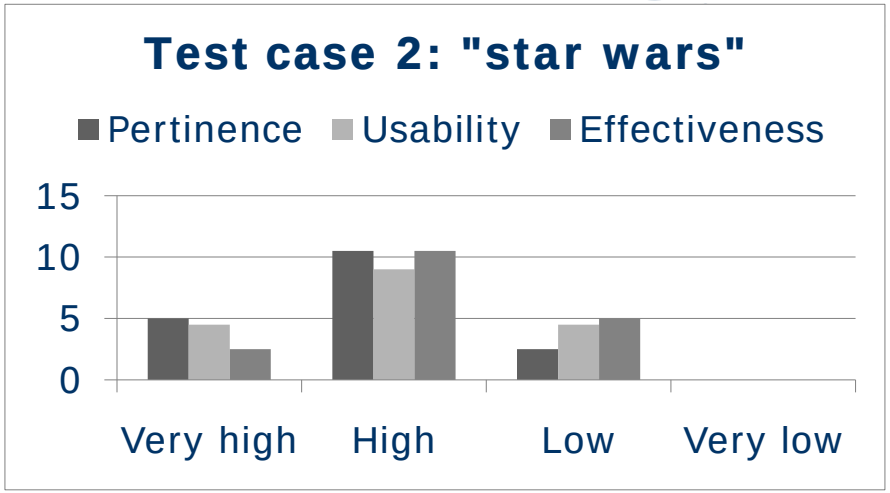
User evaluation

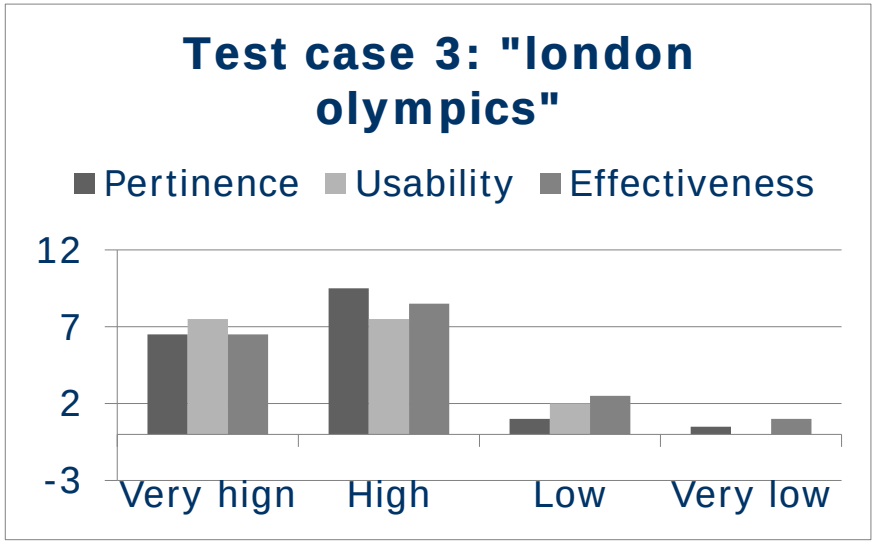
- We submitted questionnaires concerning:
  1. Pertinence of *in*Cloud organization with respect to a given target of interest
  2. Usability of *in*Cloud with respect to conventional linked data interfaces for a given target
  3. Effectiveness of *in*Clouds as a tool for the exploration of a target

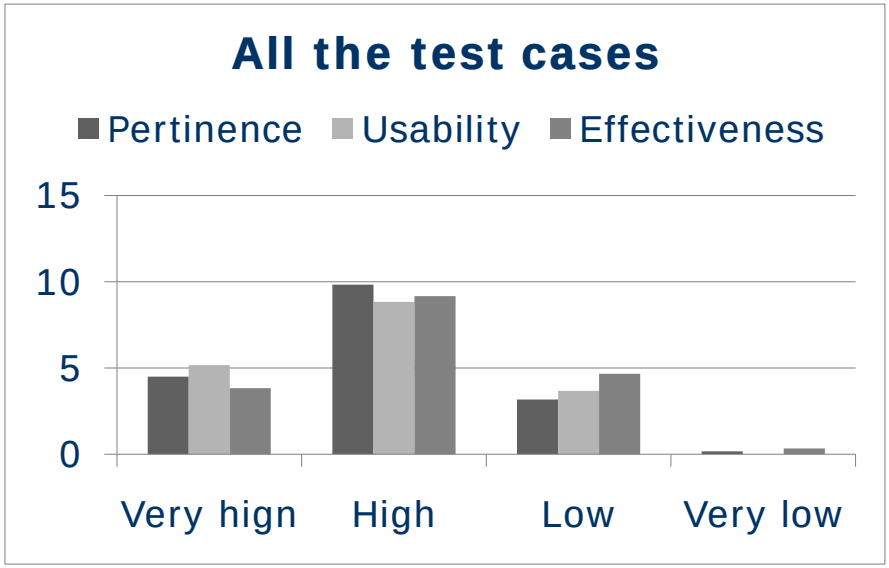
Results

IS Lab









Concluding remarks

- Preliminary results are promising about the pertinence and usability of *in*Clouds
- For future work, we would like to propose *in*Clouds as a comprehensive exploration tool considering also actual, up-to-date social web information about the search target for possible fruition in the framework of event-promoting applications
- To this end, we are mainly working on:
  - Combining more clustering algorithms in a dynamic procedure for selecting the most appropriate one to the aggregation task at hand
  - Combining data acquired from linked data and social web (i.e., microblogging, RSS)
  - Finalizing the prototype with a usable web interface

Thank you!

- More details about the project and the authors are available at

http://islab.dico.unimi.it