LIG@ALICIA

Sihem Amer-Yahia (DR CNRS@LIG)
Bilyana Taneva (post-doc ALICIA)

Also involved: **Eric Gaussier** (Prof. LIG) , **Vincent Leroy** (MdC LIG)

Kick-off meeting 20/02/2014

Outline

- **SLIDE and AMA teams @ LIG**
- **ALICIA workpackages**
- **Some technical material**

SLIDE

scalable models and algorithms for the discovery and exploitation of information and knowledge from data

- **Data acquisition and enrichment**
  - Big data preparation
  - Web data linkage
  - Crowd data sourcing

- **Large-scale data analytics**
  - Advanced pattern mining
  - Distributed join algorithms
  - Social media and health analytics

- **Information exploration**
  - Ontology-based data access
  - Interactive pattern exploration

AMA

Analyse de données, Modélisation et Apprentissage automatique

- **Data analysis and learning theories**
  - Metric learning, clustering, classification
    - Text, time series, graphs
    - Co-clustering, multiview
  - Learning theory (non i.i.d. data, large-scale, …)

- **Learning and perception systems**
  - Multi-modal models for human and robot activities
  - Self-adaptive models
  - Data fusion and uncertainty modeling
  - Human machine interaction

- **Modeling social networks**
  - Models of evolutive content networks (info. diffusion, buzz and link prediction)
  - Collective properties of social systems

Tasks we are involved in in ALICIA

- Task 1: Data Models for User-Centric Applications
- Task 2: Algorithms for Adaptive Learning
- **Task 3: Scalable Algorithms for Community Detection, Clustering and Matching**
- Task 4: User-Centric Applications
- Task 5: Evaluation

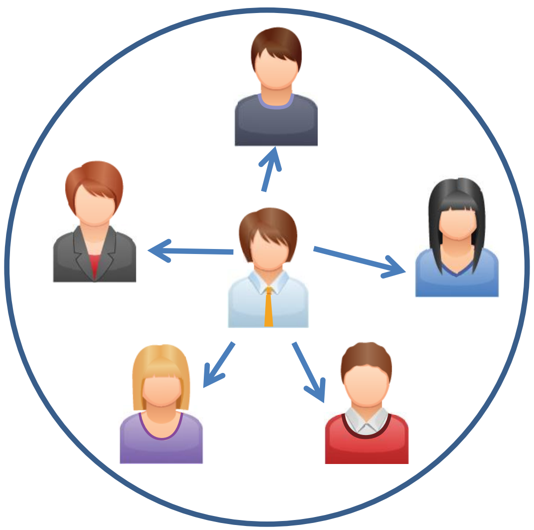
Collaborative Item/User Composition (CIC and CUC)

- Paradigm shift from *atomic* items/users to *composite* items/users
- Design a data sourcing platform for CIC/CUC where workers collaborate to produce composite items/users
- Proposed applications (task 4)
  - Search and recommendation: shift from a ranked list of items to a collection of complementary items (CIC)
  - Crowdsourcing: on-the-fly team building to achieve a task (CUC)
  - Targeted advertising: shift from single-user ad serving to a set of users (CUC)
- Leverage worker collaboration
  - *Implicit* user actions from logs
  - *Explicit* user interactions via crowd data sourcing
  - Implicit and explicit collaboration could co-exist
- Need for scalable algorithms for community detection, clustering and matching (task 3)

Examples of CIs and CUs

Vodkaster: CI = movie recommendations grouped on user preferences

CI



CU

AlephD: CU = a group of friends of 

Composite Item Retrieval (CIR)

Composite Retrieval of Diverse and Complementary Bundles @ TKDE 2014

Sihem Amer-Yahia, Francesco Bonchi, Carlos Castillo,
Esteban Feuerstein, Isabel Mendez-Diaz, and Paula Zabala

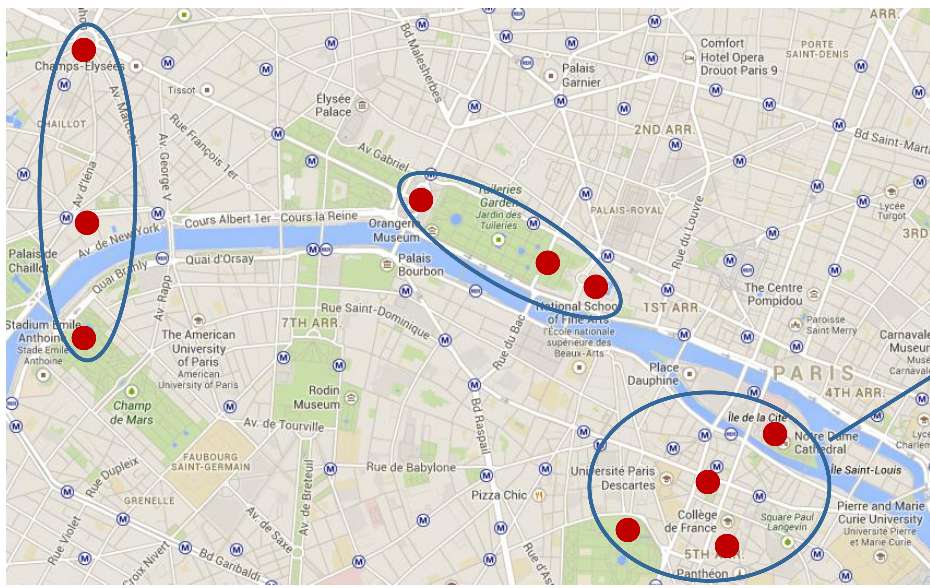
- Address complex search tasks
  - Trip planning, team building, …

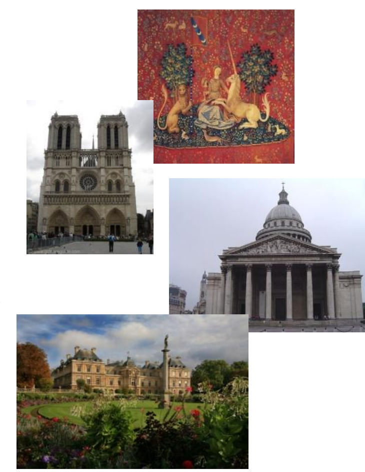
ranked list of items

1) ●
2) ●
3) ●
… ●
●
●
●
●

vs.

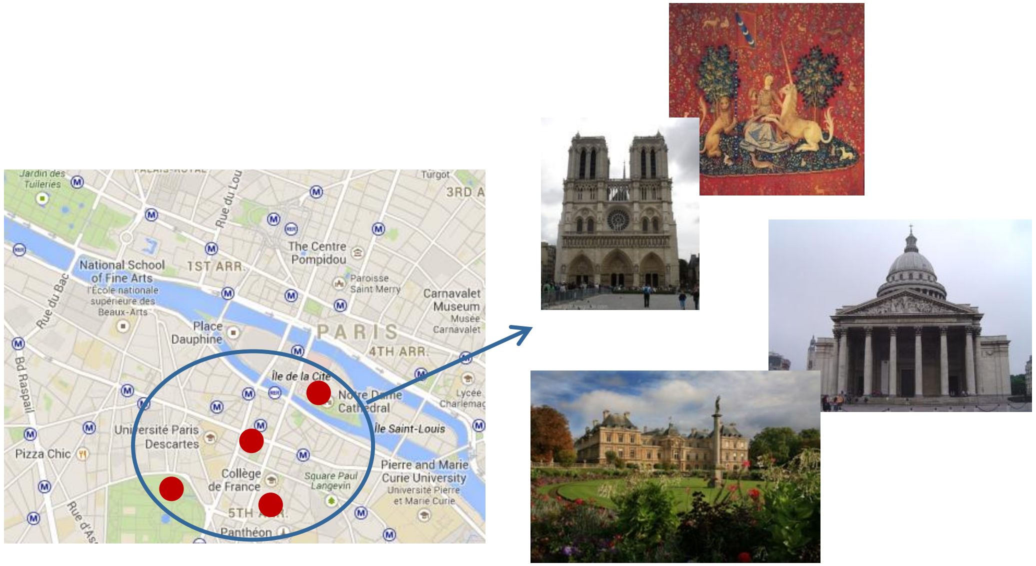
composite items





What is a Composite Item?

- A set I of (atomic) items
- Composite item: $S \in 2^I$ that satisfies
  - Complementarity: $\forall u, v \in S, u.\alpha \neq v.\alpha$
  - Budget: $f(S) \leq \beta$



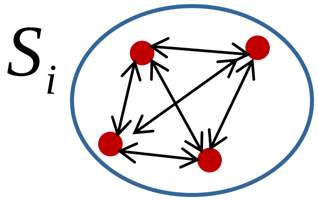
A CIR Problem

(to be adapted to applications in ALICIA)

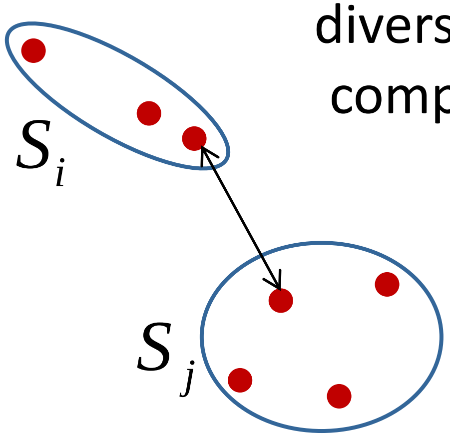
- Build (on-the-fly) a set of *k* composite items $\{S_1, \ldots, S_k\}$ that maximizes

$$\sum_i \sum_{u,v \in S_i} \gamma * comp(u,v) + \sum_{i<j} (1-\gamma) * (1 - \max_{x \in S_i, y \in S_j} comp(x,y))$$

compatibility of items in each composite item



diversity between composite items



Item Compatibility Examples

- For a pair of items *u* and *v: comp(u,v)*
- Depends on application semantics
  - Trip planning: geo proximity
  - Movie recommendations: fraction of users who like both movies, *u* and *v*
  - Ads for a group of users: friendship compatibility or other shared patterns (e.g., geo location)

CIR is hard

- CIR is NP-hard
- Two reductions of Maximum Edge Subgraph

- Each CI has only one item
  - Same complementarity value to all items

- Find only one CI
  - Different complementarity values to all items

CIR heuristics

- Produce-and-Choose
  - Two-phase approach
  - First: Produce many CIs
    - Hierarchical clustering
    - Construct CIs around randomly chosen pivots
  - Second: Choose *k* CIs
    - Adapt heuristics for the Maximum Edge Subgraph problem

- Cluster-and-Pick
  - Find *k*-clustering of all items
  - Pick the best CI from each cluster

Some insights from experiments

- Datasets
  - 20 touristic attractions in 10 cities
  - Sample of Yahoo! Local with restaurant reviews

- Compatibility vs. diversity

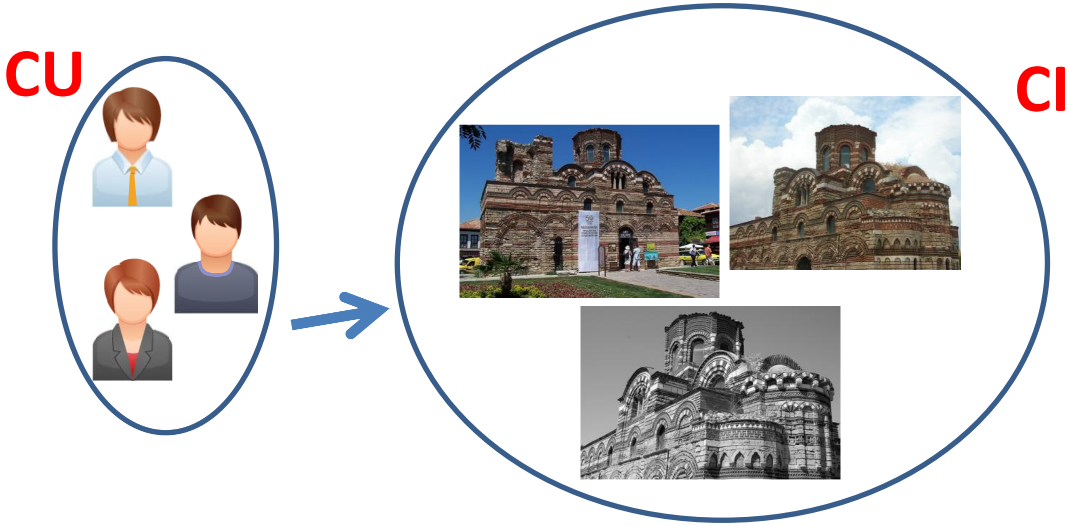
$$\sum_{i} \sum_{u,v \in S_i} \gamma * comp(u,v) + \sum_{i<j} (1-\gamma) * (1 - \max_{x \in S_i, y \in S_j} comp(x,y))$$

  - Compatibility → Produce-and-Choose
  - Diversity → Cluster-and-Pick

CIR -> CIC and CUC

- *Crowdsourced composition* relies on *explicit* user involvement for *simultaneous evaluation*

- Pinterest: form a team of users (CU) to build complementary photos of the same historical monument under different light conditions (CI)



TODO List

- A data model for CIs and CUs to represent worker interactions (task 1)
- Real data is very important to us
  – To build rich user profiles
  – For evaluation (task 5)
- Adaptive CIC and CUC algorithms (task 2)
  – Account for user interactions, feedback, actions, …
  – Account for changes in users' interests