

# Profile Diversity for Query Processing using Users Recommendations

Esther.Pacitti@lirmm.fr

Inria&Lirmm, Univeristy of Montpellier 2

Music Workshop – 20/10/2015



# Outline

1. Recommendation for Scientific Data
  - Use case: botanic data (Pl@ntnet)
2. Profile Diversity
3. Distributed and Diversified Recommendation
4. Demo
5. Conclusion

# On-Line Communities: Citizen Sciences Context\*

- Accurate knowledge of plant identities, in different geographic localities is essential for:
  - agricultural development
  - plant diversity preservation
- Pl@ntnet project\*:
  - Interactive plant identification and
  - Collaborative Information Systems
- Big Data production



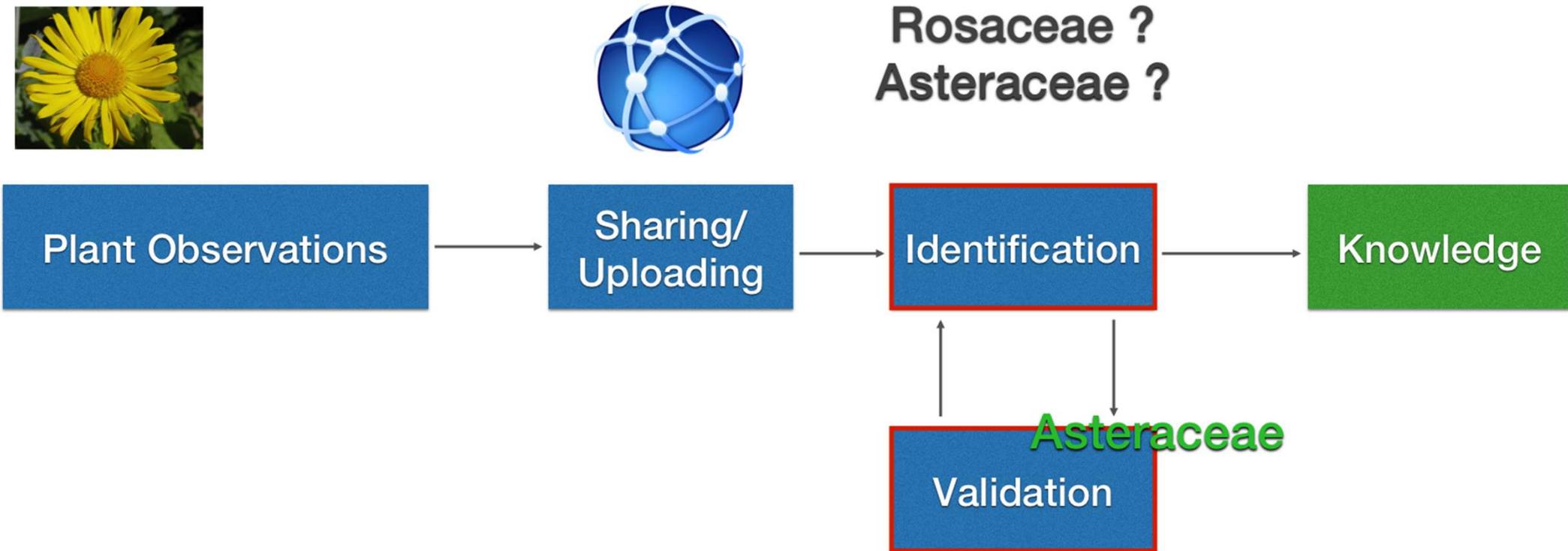
\*<http://www.plantnet-project.org/page:projet?langue=en>

# Context

- Large scale of users with different goals, profiles and interests
- Querying such data requires careful method to be able to gain access to useful information.
- Ex: **Grapevine plants identification and preservation**, requires considerable knowledge of the potential of different grape varieties and their specific morphological characters



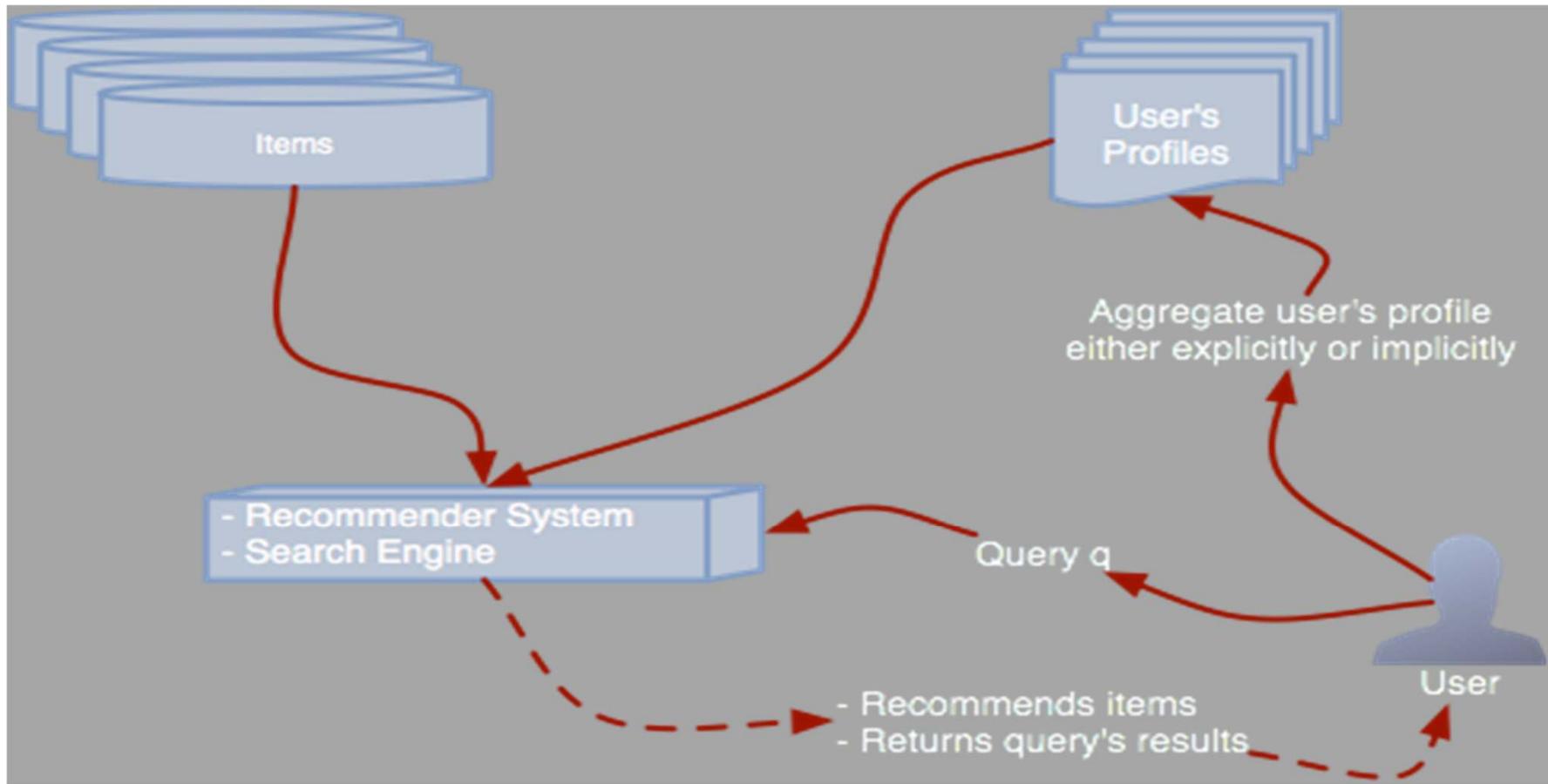
# Plant Observation Process



## General Problem:

Retrieve/recommend the  $k$  most diverse plant observations given a query (e.g. grapewine) ?

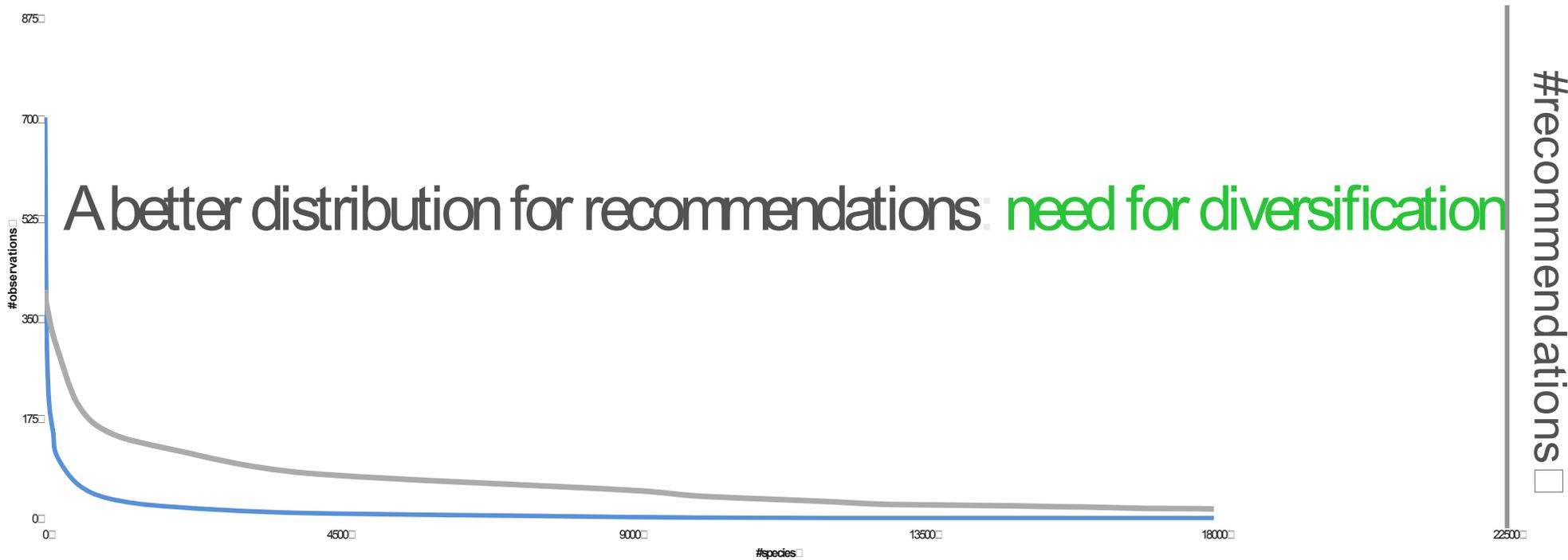
# Search and Recommendation



*Seek to recommend items given a query*

However typically recommendation methods tend to propose redundant or too popular items, because normally the **relevance** of an item wrt to query is based on **similarity or popularity**.

# Need for Diversification



A few plants represents the majority of the observations

The majority of the plants are rarely observed

[1] JOLY, Alexis, GOËAU, Hervé, BONNET, Pierre, et al. Interactive plant identification based on social image data. Ecological Informatics, 2013.

[2] <http://www.bugwood.org>, 2014

# Diversification in Search in Search and Recommendation

- Recall in information retrieval is the fraction of the items that are relevant to the query that are successfully retrieved.
- Given query  $q$  and the candidates items for  $q$ , the diversity is measured by measuring the distance of among selected *items* in the response.



# Use Case with Delicious

Methods	Case 1 No Diversity	Case 2 Content Diversity
Users	User 1	User 2
Profiles	Sea, boat, fishing	Savanna, jungle, Africa
Results		

**Content Diversity** [Angel, Sigmod 2011] has been used with promising results. However, it presents low diversity gains due to:

- poor content description
- semantic ambiguity

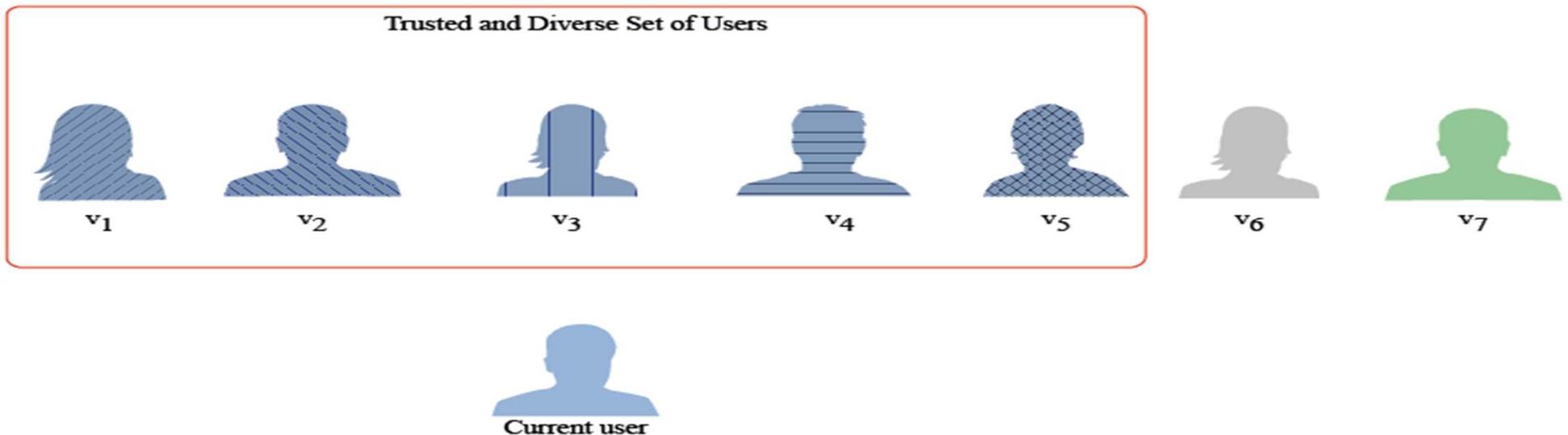
Ex: [Java Language](#) x [Java Island](#)

# Profile Diversity\*



- Users profiles are defined based on the items they share
- The relevancy takes into account **similar and diversified users**, wrt to the items they share.

*With **profile diversity\***, the recommended items, in addition to be relevant to the query, and to the user profile, also take into account the diverse relevant users profiles and their items.*



# Search and Recommendation

$q = \text{asteraceae}$		
Un-diversified	Profile Diversity	
Élodie Dujardin	Marie Dupont	Pierre Durand
		
		
		

Table I: Search and recommendation example with profile diversity.

ASTERACEAE, *Leucanthemum adustum* - Montpellier (34)  
 ASTERACEAE, *Leucanthemum atratum* - Montpellier (34)  
 ASTERACEAE, *Cichorium intybus* - Montpellier (34)  
 ...

(a) Élodie's Profile

ASTERACEAE, *Cirsium vulgare* - Palavas (34)  
 ASTERACEAE, *Cichorium intybus* - Paris (75)  
 ASTERACEAE, *Crepis vesicaria* - Montpellier (34)  
 ...

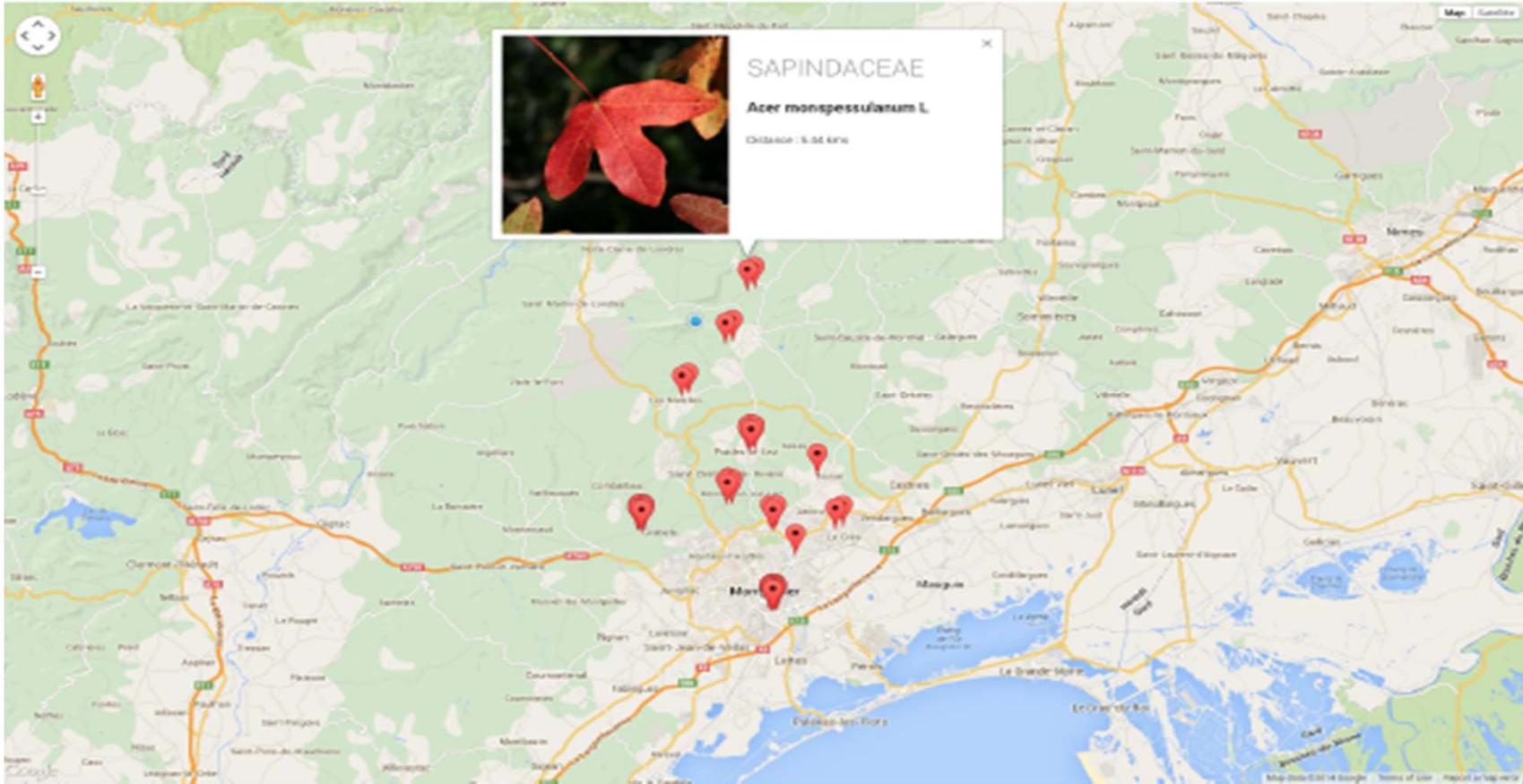
(b) Marie's profile.

RHAMNACEAE, *Ramnus alaternus* L. - Montpellier (34)  
 ASTERACEAE, *Echinops ritro* L. - Montpellier (34)  
 ASTERACEAE, *Senecio doronicum* - Montpellier (34)  
 ...

(c) Pierre's profile.

Figure 2: User profiles.

# Diversity of Plants in a Geographical Locality



Provide all plant diversity given a geographical area (bleu point)

# Search and Recommendation Model

$$\text{score}(it_i, u, q) = \text{rel}(it, q) \times \text{div}_c(it_i | \{it_1, \dots, it_{i-1}\}) \times \text{div}_p(u_{it_i} | u_{\{it_1, \dots, it_{i-1}\}})$$

Relevance

Content Diversity

Profile Diversity



$$\text{div}_p(u_{it_i} | u_{\{it_1, \dots, it_{i-1}\}}) = \frac{1}{N} \times \sum_{v_n \in u_{it_i}} \text{rel}(u, v_n, q) \times \prod_{v_m \in u_{\{it_1, \dots, it_{i-1}\}}} 1 - \text{red}(v_n | v_m)$$

Profile Diversity

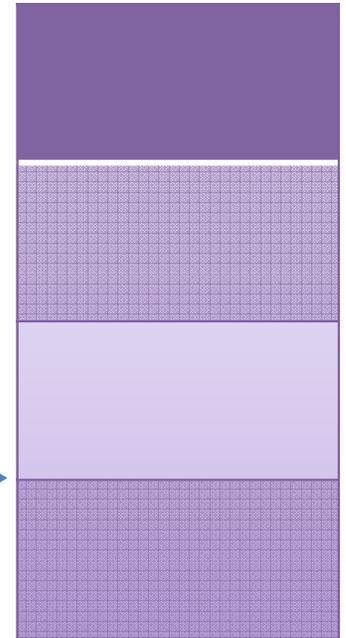
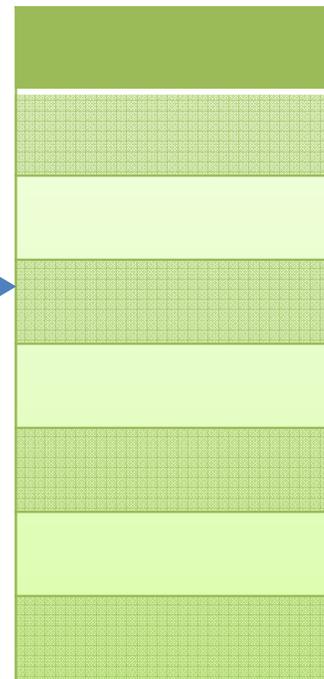
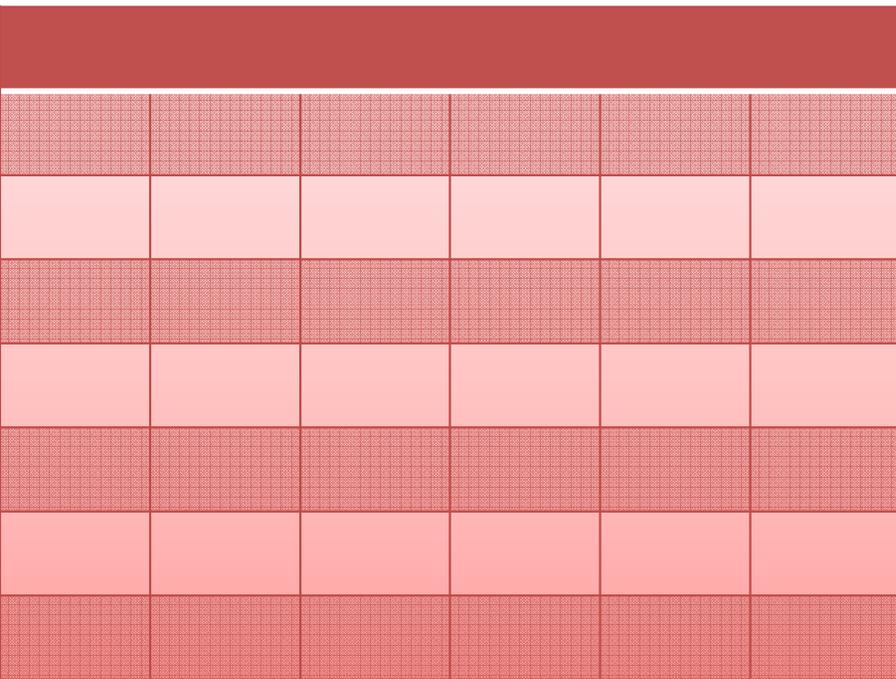
User Relevance

User Diversity

**Good Compromise between Relevancy and Diversity**

# Search and Diversified Recommendation

**Inverted Lists (Items,Users)**

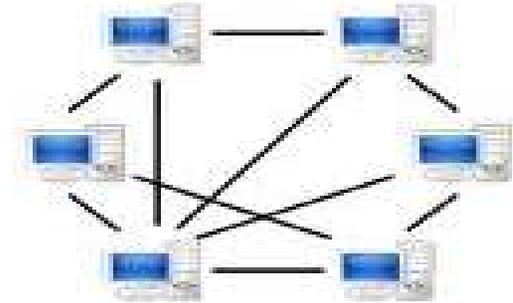


**Result List (Items)**

**Candidates List (Items, Users)**

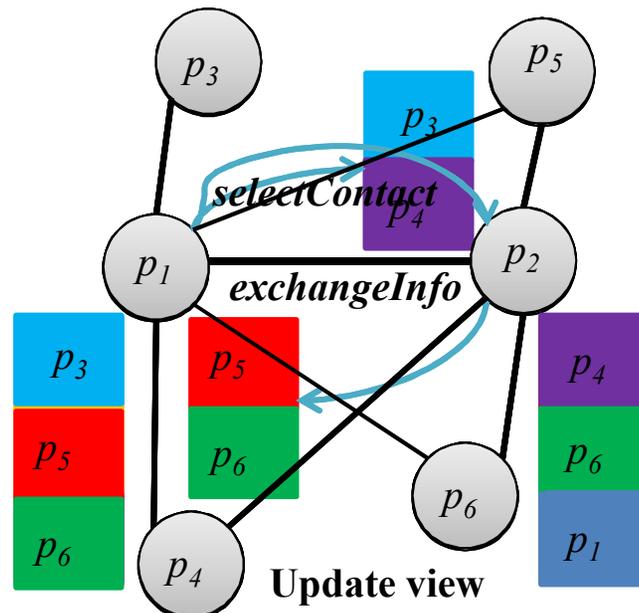
# Distributed Recommendation for Citizen Sciences

- Single users
  - Users keeps observations in their own workspace
  - Scales-up (more network traffic)
  - Better for privacy
- Multi-Sites center (local clouds)
  - Users keeps observations in the local site server
- Hybrid



# Distributed Search and Recommendation Model

- **Model:**  $G = (U, I, E)$ ,  $U$  = users nodes,  $I$  = items,  $E$  = edges among users
- **User profiles** are defined based on the items they share
- **User network (U-net):** refers to the cluster of relevant users profiles a users  $u$  is aware of through epidemic protocols.
- An edge exists between a  $u$  and  $v$ , if  $v$  is in  $u$ 's U-net.
- When a **keyword query** is submitted, it is recursively redirected to the **Top-n** similar users in the **U-net**, until TTL.
  - Each involved user computes recommend the most relevant items.



# Diversity to increase Coverage

- Coverage
  - probability of finding relevant users to provide relevant items for a given query
  - depends on the clustering metric
- Clustering metric
  - similarity (e.g. cosinus, jaccard, etc).[Draidi 10, Xiao 2011, Isaila 12, Kermarrec 12]
  - at each exchange  $p_l$  discovers new users and computes:  
 $\text{similarity}(\text{profile}(p_l), \text{new profiles}(p_i))$   
 $p_l$  keeps only the most similar users  $p_i$  in its U-net
- **Problem:** Recall results are low, because a significant amount similar users are kept in similar users U-nets. How to increase the quality of the coverage ?

# Clustering Useful Users\*

What if we exploit **usefulness instead of similarity** ?

- The usefulness of new user profile is determined based on the set of users profiles previously clustered in U-Net.

At each epidemic, exchange,  $u_l$  discovers new peers and computes

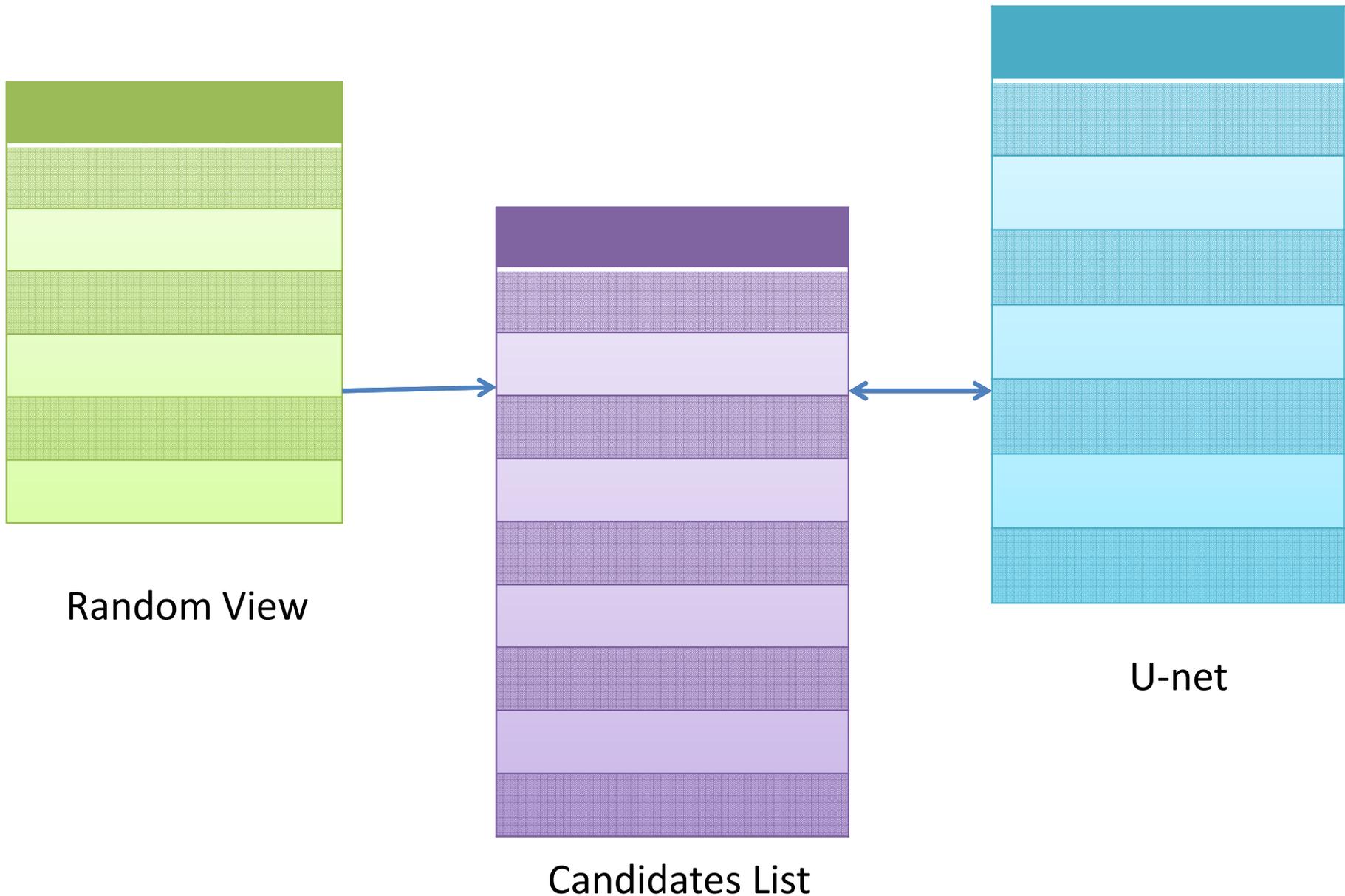
**usefulness(U-net, profile( $u_l$ ), new profiles( $u_j$ )),**

and keep only **useful** peers with best scores in  $u_i$  in **U-net**.

$$usefulness(v_j | v_{j+1}, \dots, v_n) = rel(v_j) \times \prod_{i \in j+1, \dots, n} (1 - red(v_j, v_i))$$

\*[Servajean et al. Globe 14]

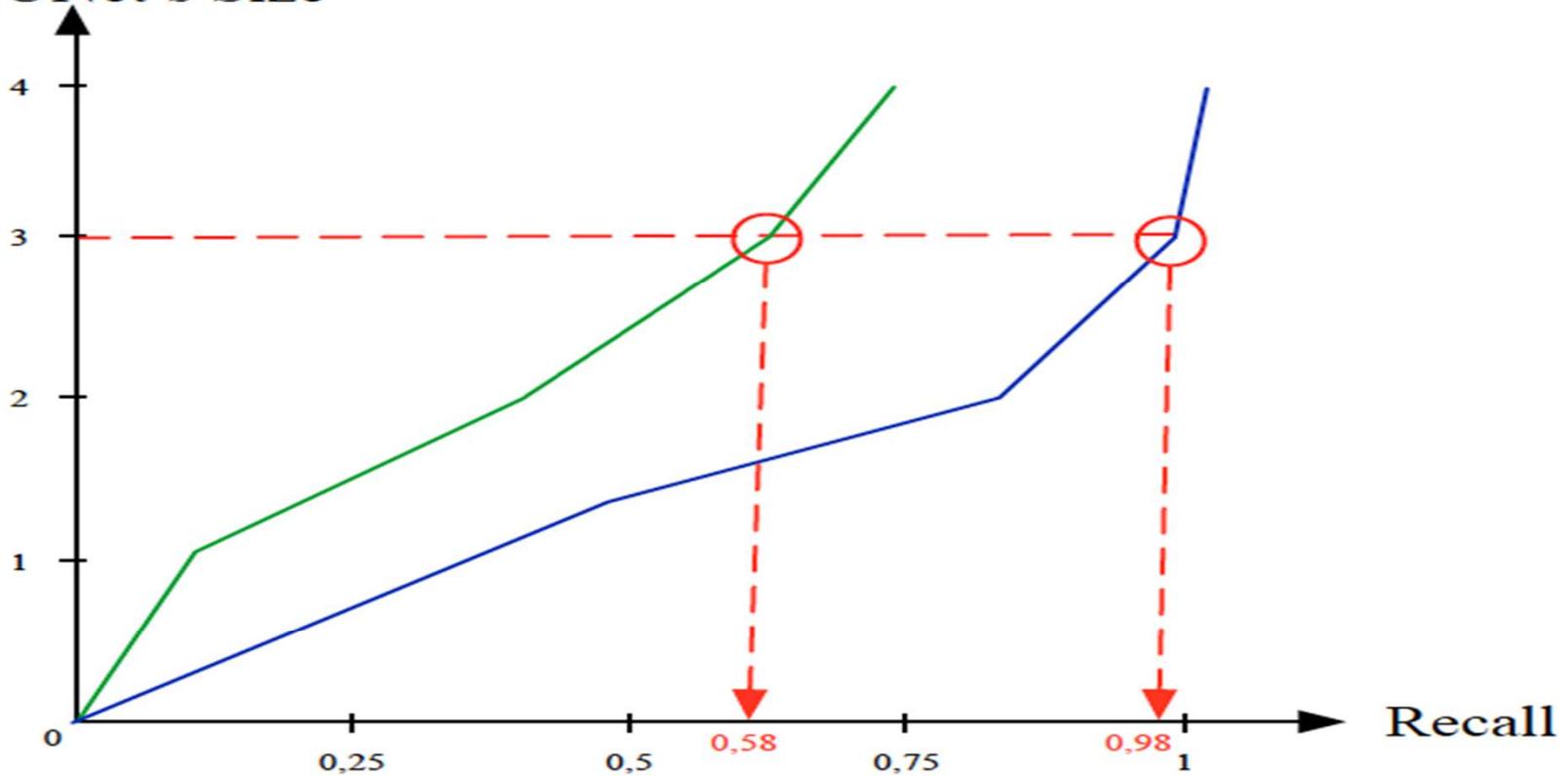
# Dynamic Clustering Algorithm for Useful Users



# Experimental Results

- Gains in Recall

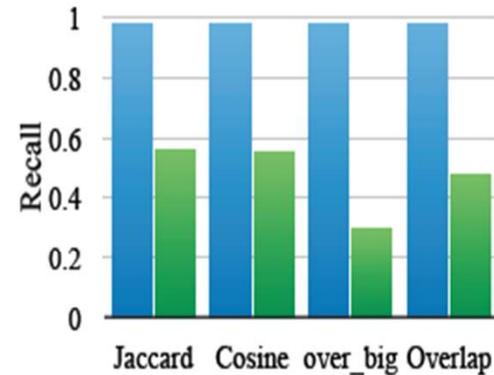
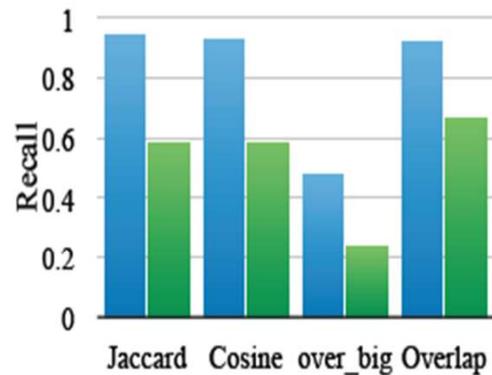
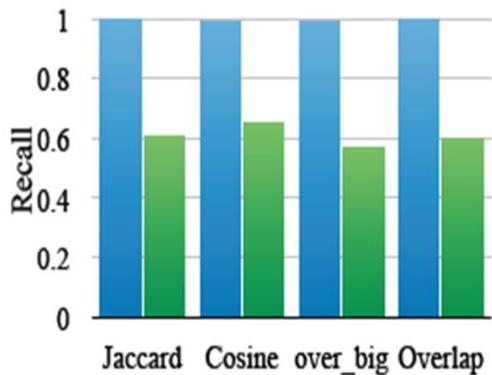
TTL / UNet's Size



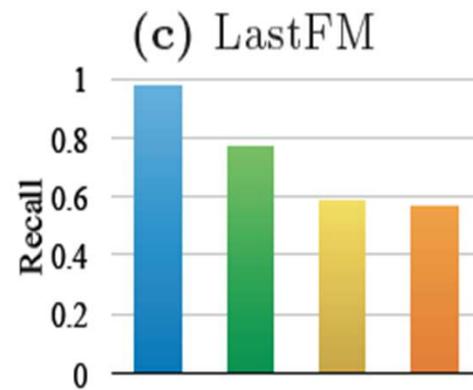
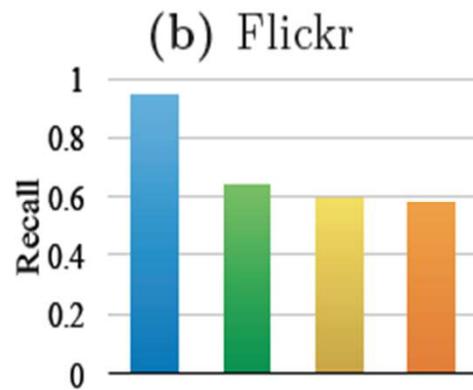
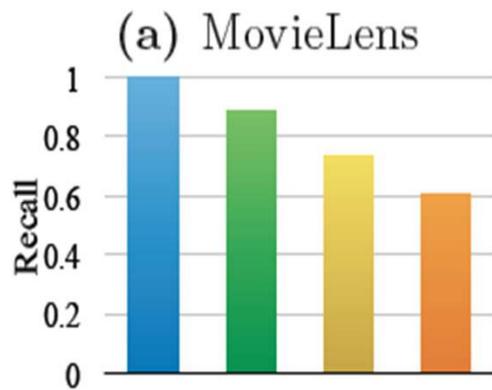
— : with diversity

— : without diversity

# Experimental Results



■ Diversified  
■ Undiversified



■ Usefulness (Jaccard)  
■ xQuad (Jaccard)  
■ MMR (Jaccard)  
■ Jaccard

(d) MovieLens

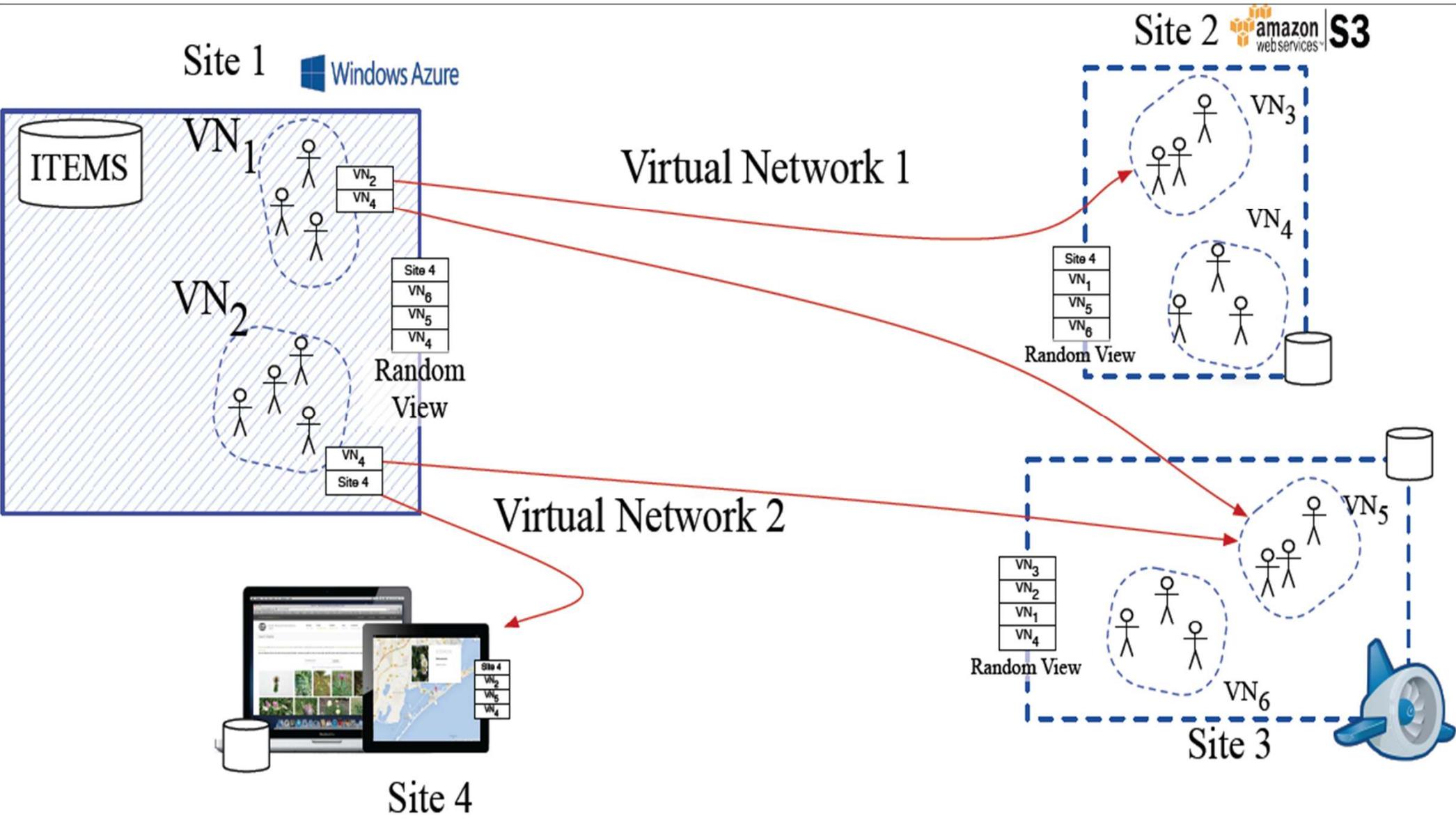
(e) Flickr

(f) LastFM

TTL = 3

For details: [Servajean et al. Globe 14]

# Diversity in Multi-Site Recommendation\*





## Plant Recommendation Tool

### Authentication

Type your username and password below

---

Forgot password ? [Click here to reset](#)

You do not have an account ? [Register](#)

---

# Conclusion