USC

USC Viterbi
School of Engineering

Growing a Tree in the Forest: Constructing Folksonomies by Integrating Structured Metadata

Anon Plangprasopchok (USC/ISI)
Kristina Lerman (USC/ISI)
Lise Getoor (UMD)

User-Generated Content & Metadata on the Web

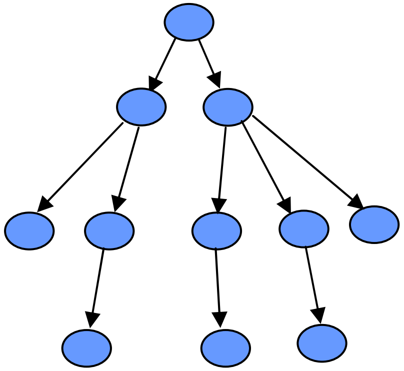
- **Explosion of user-generated content**
  - ***Images: Flickr, Picasa,..***
  - ***Videos: YouTube, Vimeo,..***
  - ***Maps: WikiMapia,..***
  - ***Story: Blogs, Twitter,..***
  - ***Relational Data: Metaweb, Google Base,..***

- **User-generated semantics: annotation/metadata**
  - ***Tags, Geotags***
  - ***Personal Hierarchies***

Goal: extract users' knowledge (folk knowledge) from this metadata

Folksonomy (communal taxonomy/hierarchy)

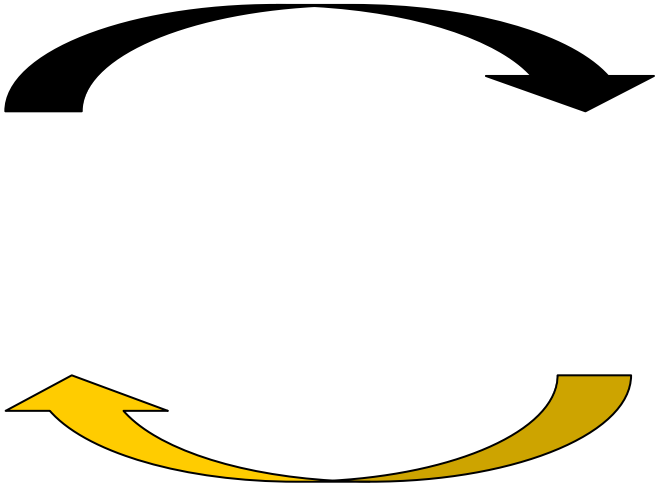
Folksonomy that users commonly have in their *mind (hidden)*



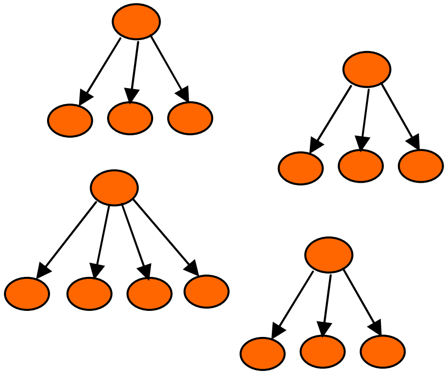
...

[deep & bushy]

Users select a portion of the hierarchy to organize their content.



Personal hierarchies from various users (*observed*) such as users' folder-sub folders



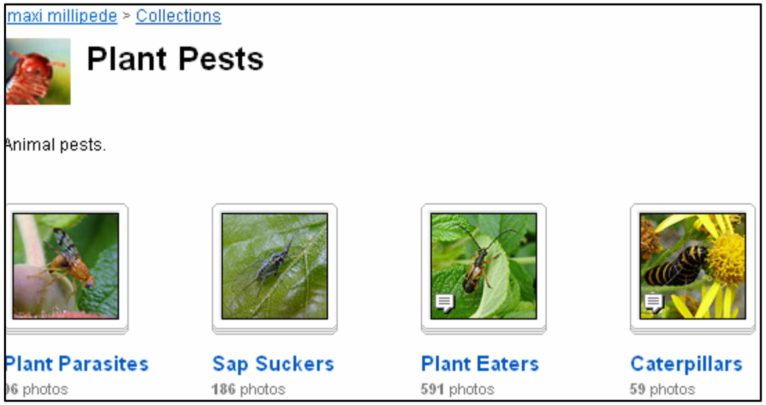
[shallow, noisy, sparse(incomplete) & inconsistent]

Can we recover the folksonomy back from many observed hierarchies? → folksonomy learning

Outline

- **Personal Hierarchies in Social Web**
- **Challenges in Folksonomy Learning**
- **Integrate Personal Hierarchies to Folksonomy**
  - ***Relational Clustering for Learning Folksonomy***
- **Evaluations**
  - ***Metrics***
  - ***Results***
- **Related Work**
- **Discussion & Conclusions**

Structured Social Metadata in Flickr



maxi millipede > Collections

Plant Pests

Animal pests.

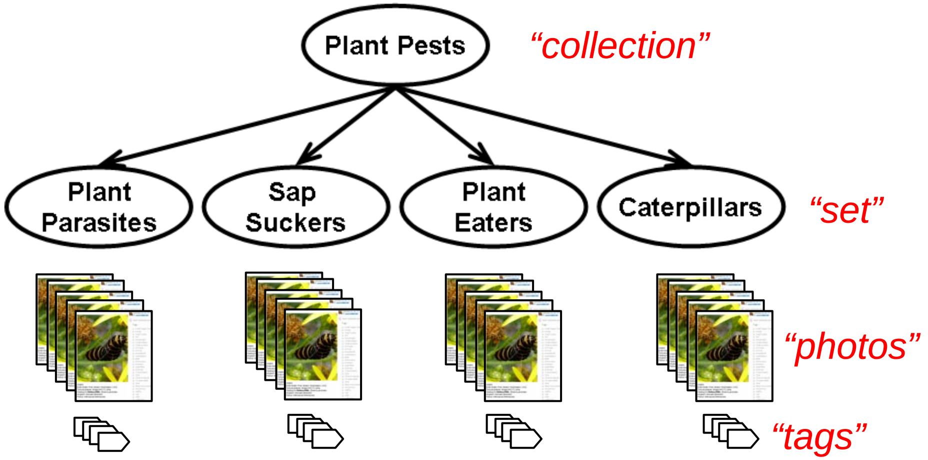
Plant Parasites
96 photos

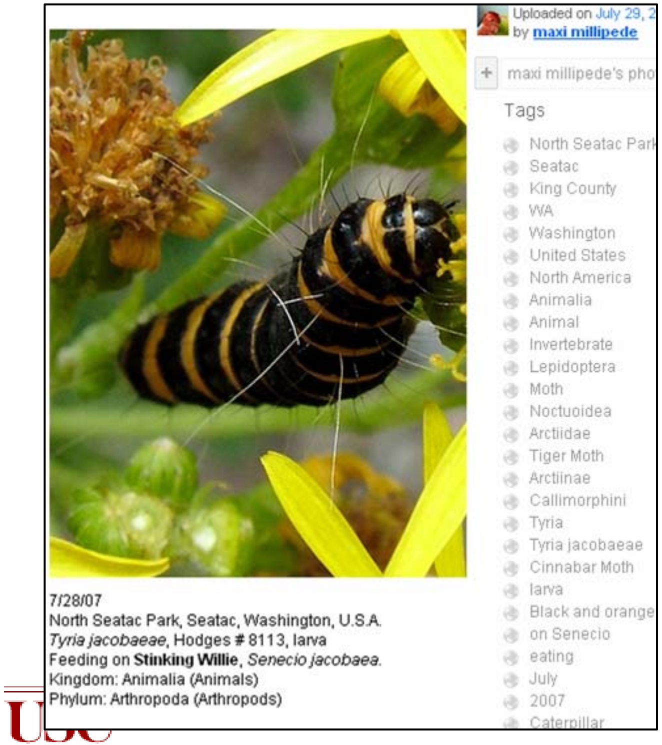
Sap Suckers
186 photos

Plant Eaters
591 photos

Caterpillars
59 photos

Shallow Hierarchy by user "maxi millipede"





Uploaded on July 29, 2
by maxi millipede

maxi millipede's pho

Tags

- North Seatac Park
- Seatac
- King County
- WA
- Washington
- United States
- North America
- Animalia
- Animal
- Invertebrate
- Lepidoptera
- Moth
- Noctuoidea
- Arctiidae
- Tiger Moth
- Arctiinae
- Callimorphini
- Tyria
- Tyria jacobaeae
- Cinnabar Moth
- larva
- Black and orange
- on Senecio
- eating
- July
- 2007
- Caterpillar

7/28/07
North Seatac Park, Seatac, Washington, U.S.A.
Tyria jacobaeae, Hodges # 8113, larva
Feeding on **Stinking Willie**, *Senecio jacobaea*.
Kingdom: Animalia (Animals)
Phylum: Arthropoda (Arthropods)

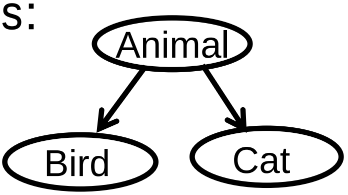
Tags on each photo

<u>Assume</u>:

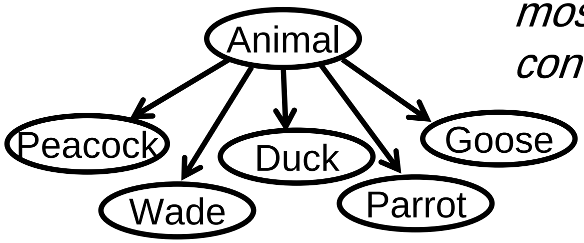
1) A set of tags on a set is an aggregation of all tags of all photos in the set
2) A set of tags on a collection is an aggregation of all tags of all sets in the collection

Challenges in Folksonomy Learning

1.) Sparseness:



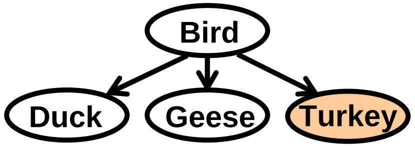
ubiquitous

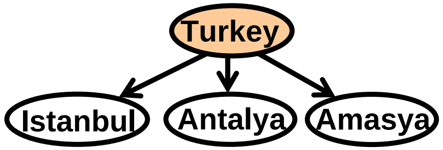


very rare!

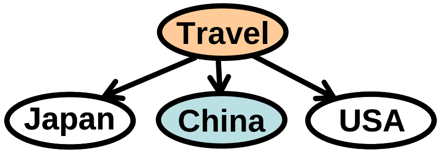
mostly personal hierarchies contain very few child nodes

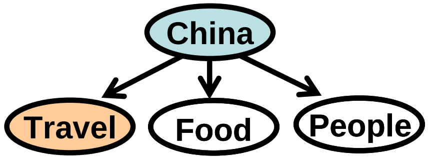
2.) Ambiguity:



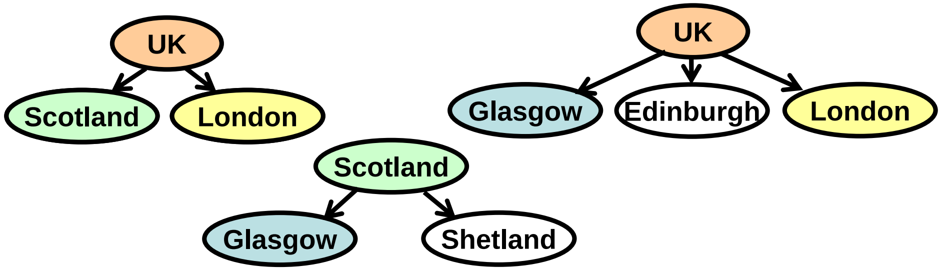


3.) Conflict:



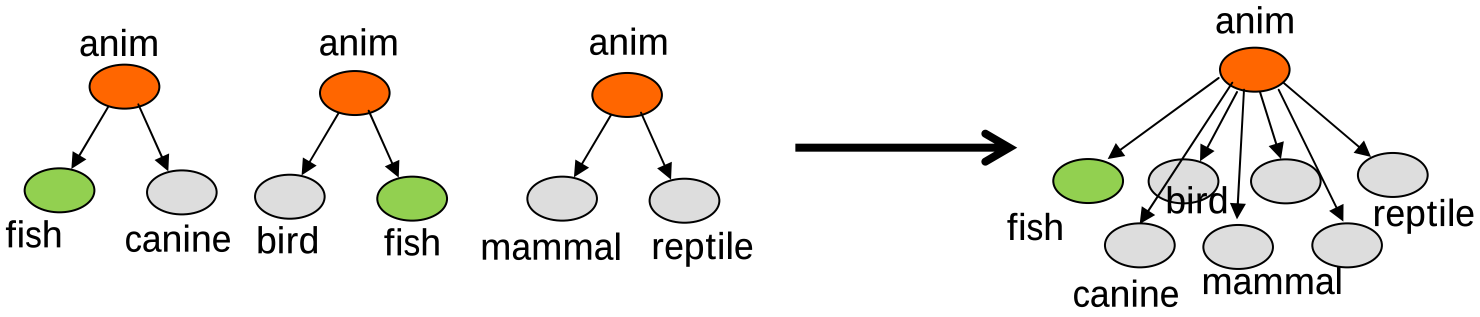


4.) Varying Granularity:

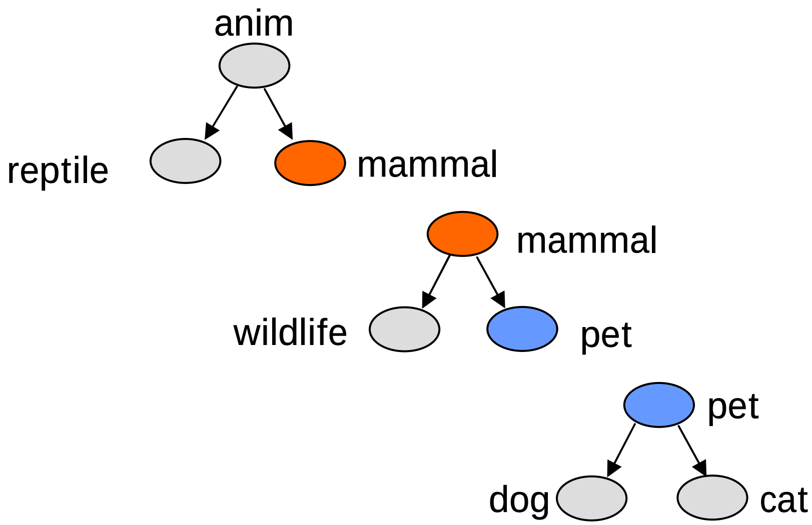


Folksonomy Learning Approach

Sketched idea: ***combine/aggregate personal hierarchies together in both horizontal and vertical directions.***



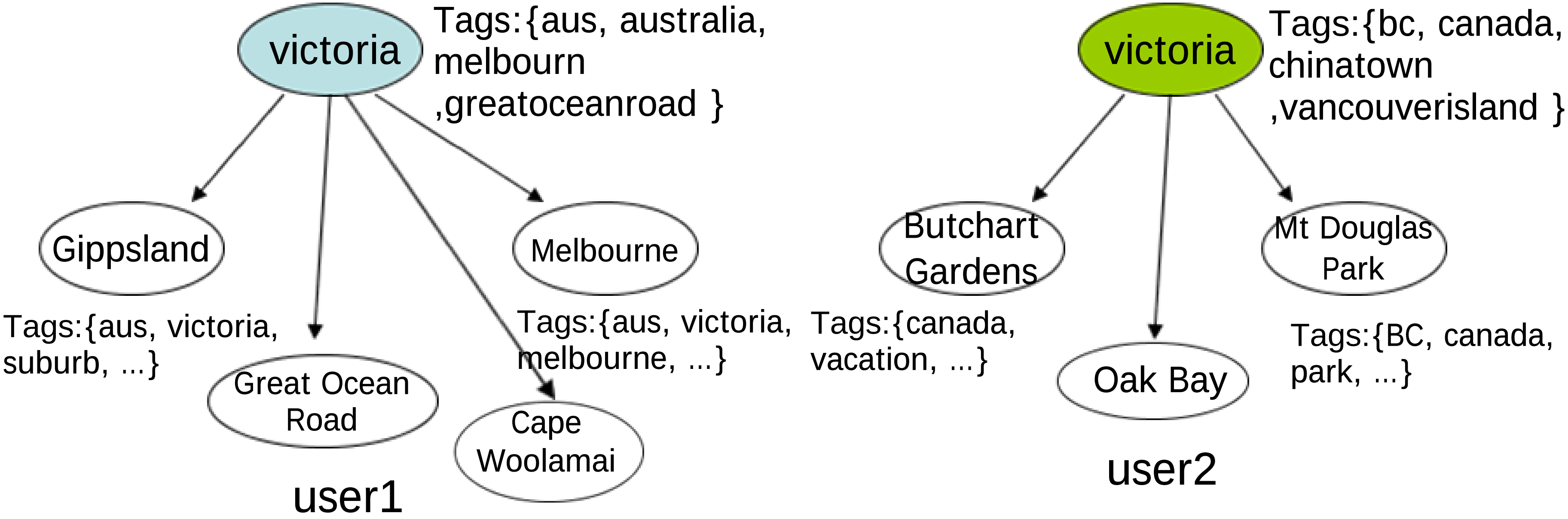
Horizontal aggregation: expanding folksonomy's width



Vertical aggregation: extending folksonomy's depth

Relational Clustering for Learning Folksonomy

Sketched idea: 2 nodes should be clustered if they are <u>similar</u> enough – similarity is computed using *contextual & relational information*



victoria

Tags:{aus, australia, melbourn ,greatoceanroad }

Gippsland

Tags:{aus, victoria, suburb, ...}

Great Ocean Road

Cape Woolamai

Melbourne

Tags:{aus, victoria, melbourne, ...}

user1

victoria

Tags:{bc, canada, chinatown ,vancouverisland }

Butchart Gardens

Tags:{canada, vacation, ...}

Oak Bay

Mt Douglas Park

Tags:{BC, canada, park, ...}

user2

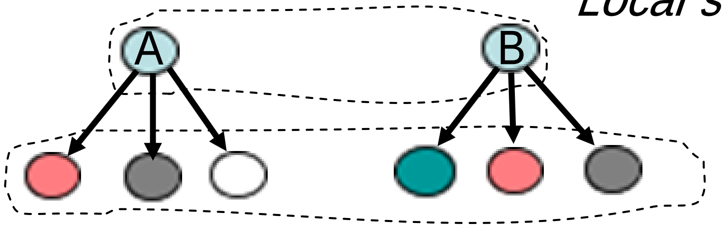
Check common tags & child nodes

Relational Clustering for Folksonomy Learning

Formally, two nodes are considered similar if:

(1) their features are similar, i.e., have similar names, have many common tags – *local similarity*

(2) their neighbors are similar – *structural similarity*



Local similarity: sim(A,B)

Structural similarity: sim(neighbors(A), neighbors(B))

$$\boxed{\mathrm{Sim}(A,B) = (1-\alpha)*\mathrm{localsim}(A,B) + \alpha*\mathrm{structuralSim}(A,B)}$$

α is a weight on how much we rely on structural information

We then merge nodes together if they are similar enough.

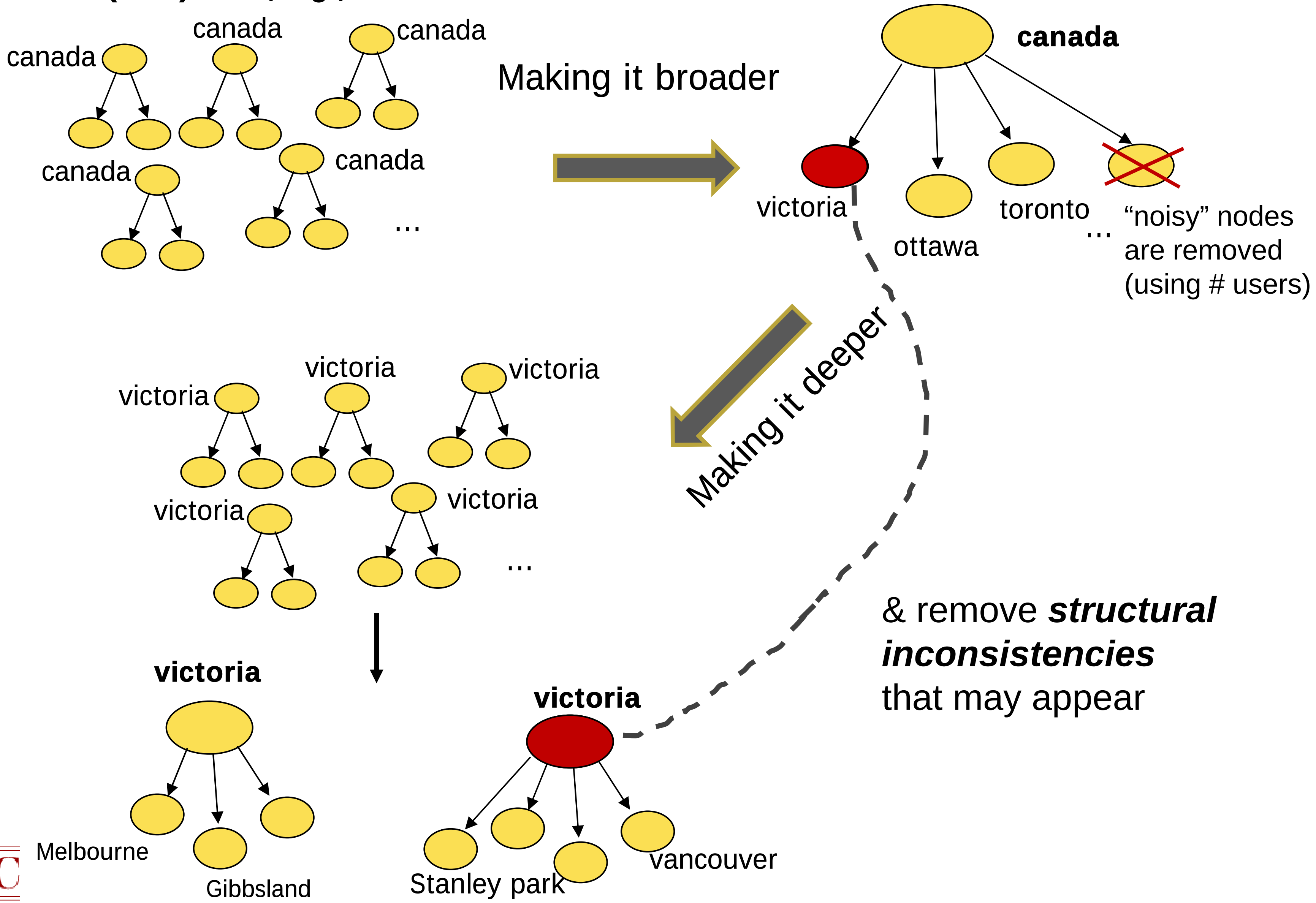
Note that we use naïve version of relational clustering by simply using neighbors' local features to compute structural similarity, instead of neighbors' class labels*.

*see Bhattacharya & Getoor, 2007, Collective Entity Resolution in Relational Data, TKDD for more detail

Incremental Relational Clustering for Learning Folksonomy

USC Viterbi School of Engineering

- **Pick a seed(root) term, e.g., “canada”**



Making it broader

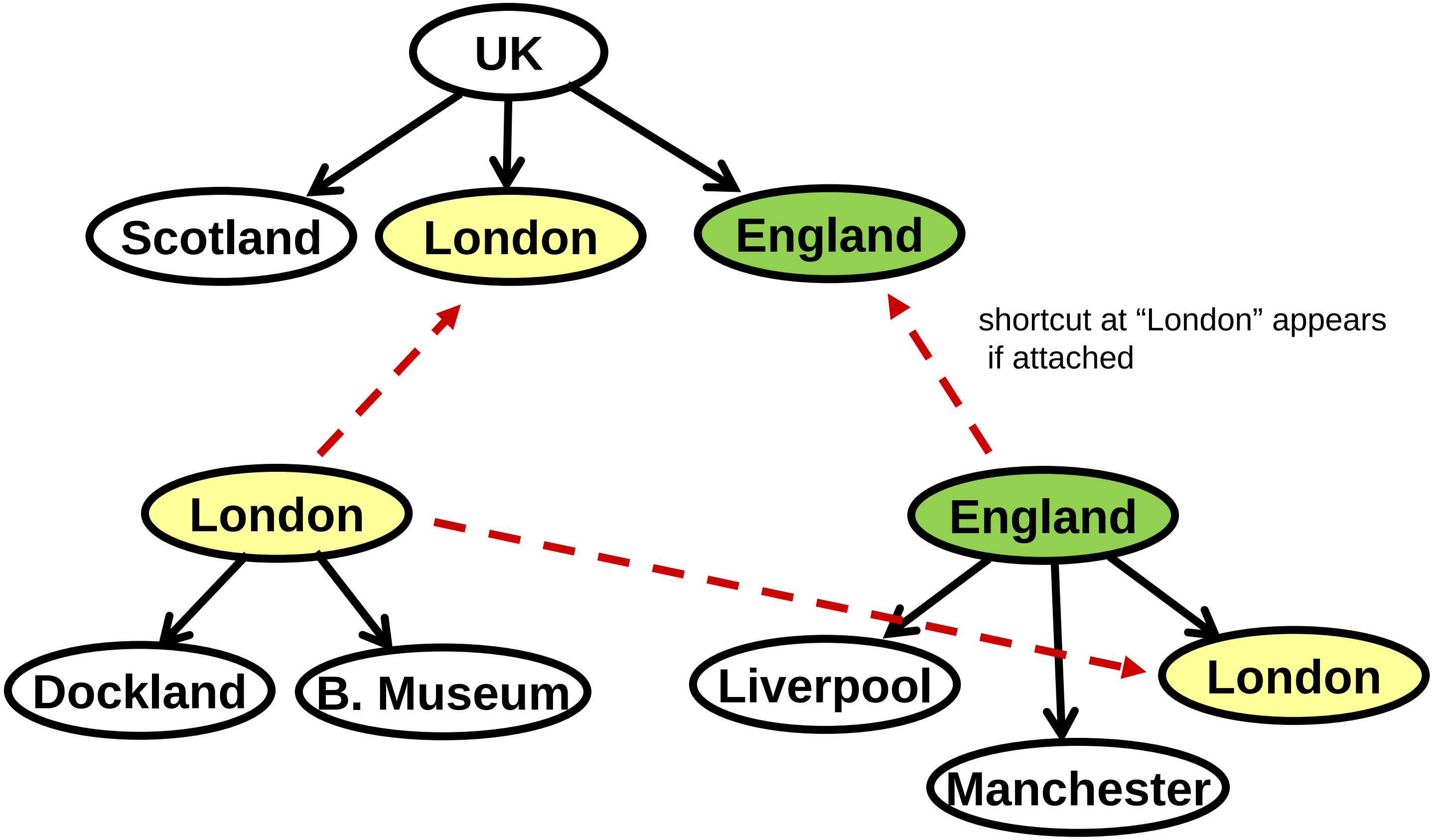
“noisy” nodes are removed (using # users)

Making it deeper

& remove ***structural inconsistencies*** that may appear

Handling Structural Inconsistencies: shortcuts

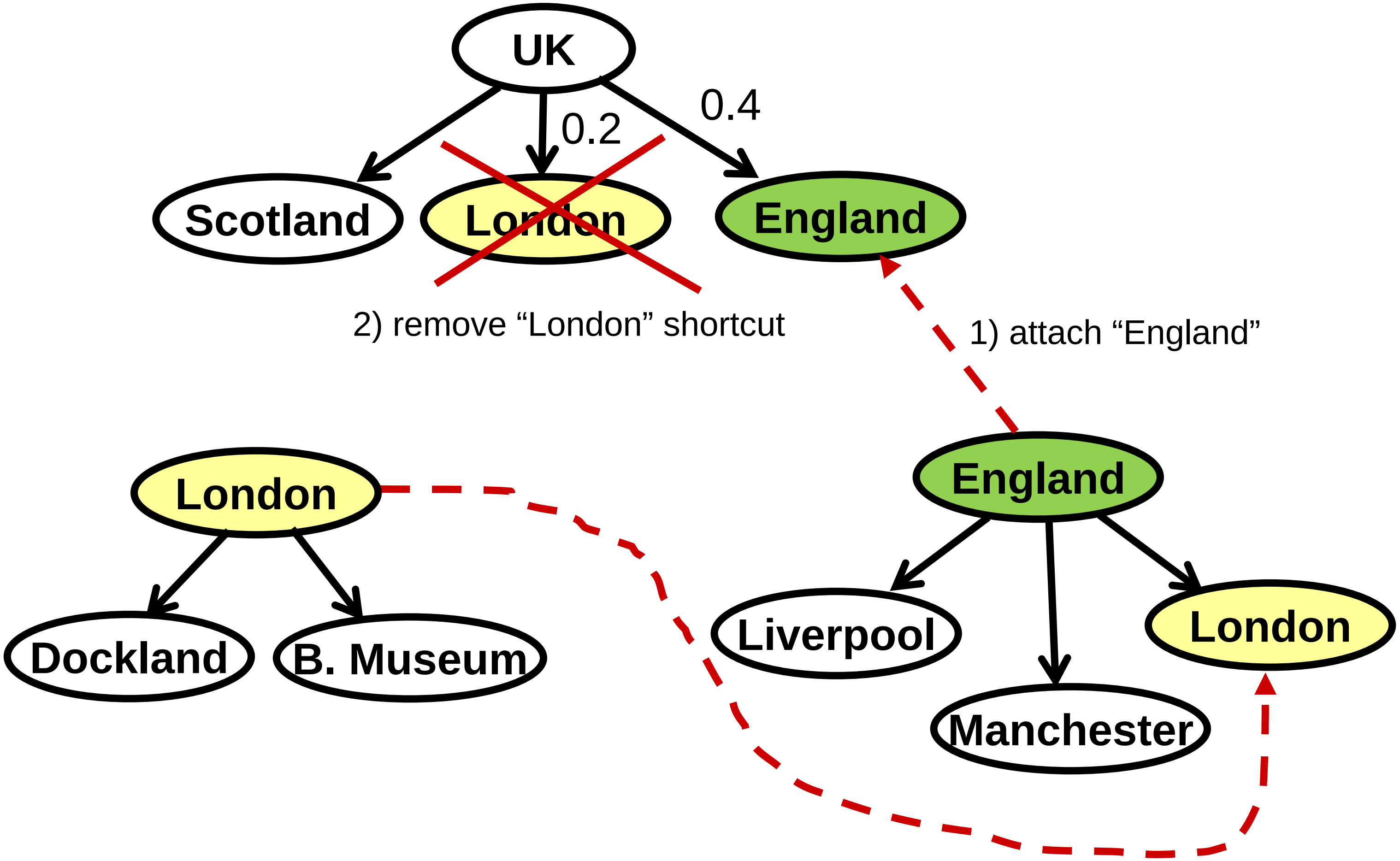
Suppose we have the following clusters of hierarchies:



shortcut at "London" appears if attached

- Shortcuts have to be removed to make the learned hierarchy consistent

Handling Structural Inconsistencies: shortcuts (2)



2) remove "London" shortcut

1) attach "England"

3) Attach "London"

- Keep the longer path since it captures more specific knowledge

Evaluations

- Growing trees from 32 seed terms & uses personal hierarchies from Flickr as in the previous work.*

Evaluation Methodologies:

1) *Against the reference hierarchy* (DMOZ)
2) *Structural evaluation*
3) *Manual evaluation*

Baseline Approach*

- Assume nodes having the same name refer to the same concept
- Keep the relations between node pairs if they are not generated at random (using significance test)
- Then, combine all relations into a tree

* A. Plangprasopchok and K. Lerman, 2009, Constructing folksonomies from user-specified relations on flickr, WWW

Evaluation Metrics

1.) an automatic comparison to the reference hierarchy

-**Taxonomic Overlap** [adapted from Maedche & Staab] measuring structure similarity between two trees. For each node, determining how many ancestor and descendant nodes overlap to those in the reference tree.

-**Lexical Recall** measuring how well an approach can discover concepts, existing in the reference hierarchy (coverage)

2.) Structural evaluation

-**Area Under Tree (AUT)** combining bushiness and depth of the tree into a single number: the higher value, the bushier and deeper tree.

3.) Manual evaluation

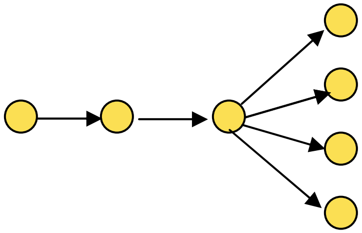
- **Accuracy**: simply asking users whether a path from root to leaf of is correct: if there are some nodes misplaced in the wrong order, users will judge the whole path incorrect

*A. Maedche & S. Staab, 2002, Measuring Similarity between Ontologies, in EKAW

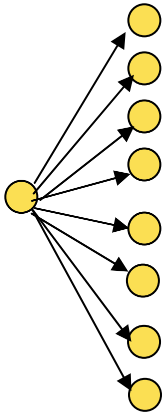
Area Under Tree (AUT)

Which structures are the best in term of “bushiness” and “depth”?

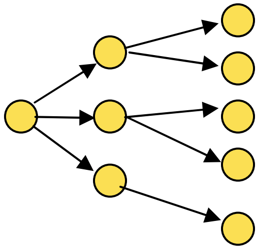
a)



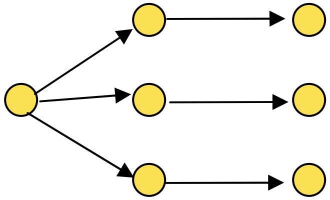
b)

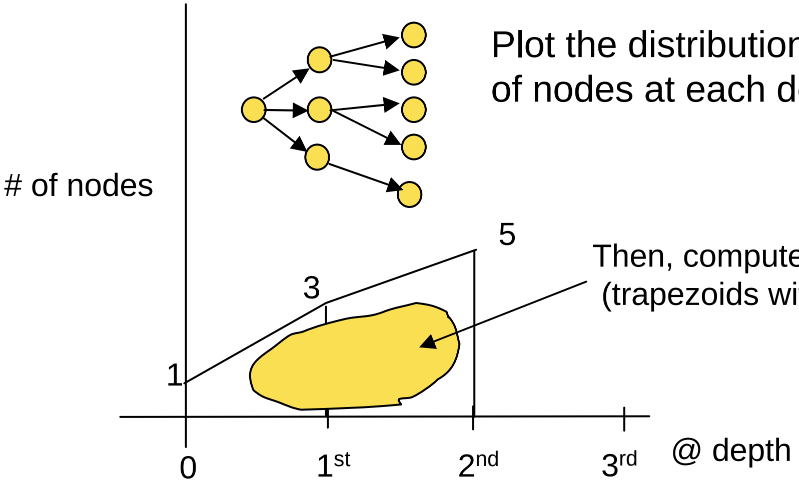


c)



d)



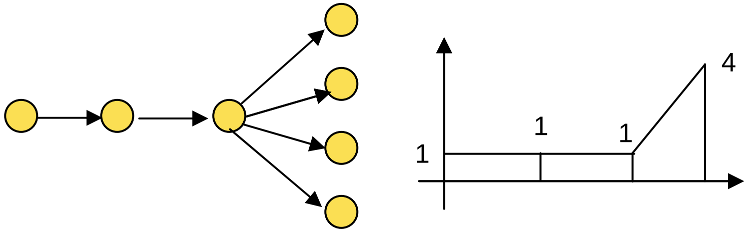


Plot the distribution on # of nodes at each depth

Then, compute the area here (trapezoids with height value = 1)

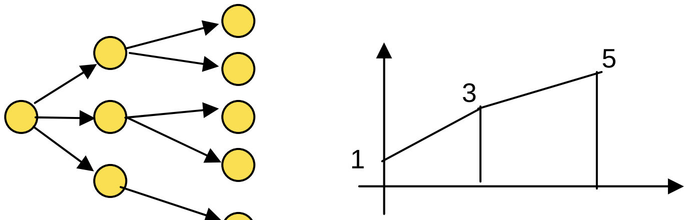
Area Under Tree (AUT)

a)



AUT = 4.5

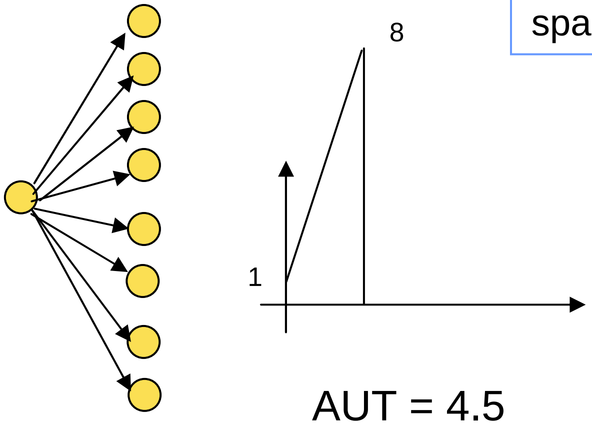
c)



AUT = 6

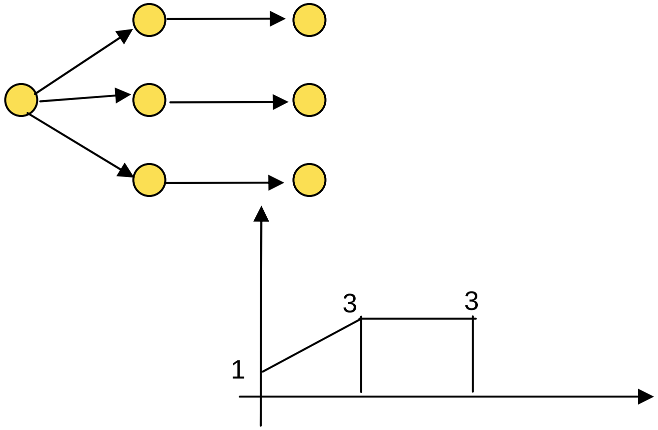
The largest area we can get is from the tree that keeps spanning at each level

b)



AUT = 4.5

d)



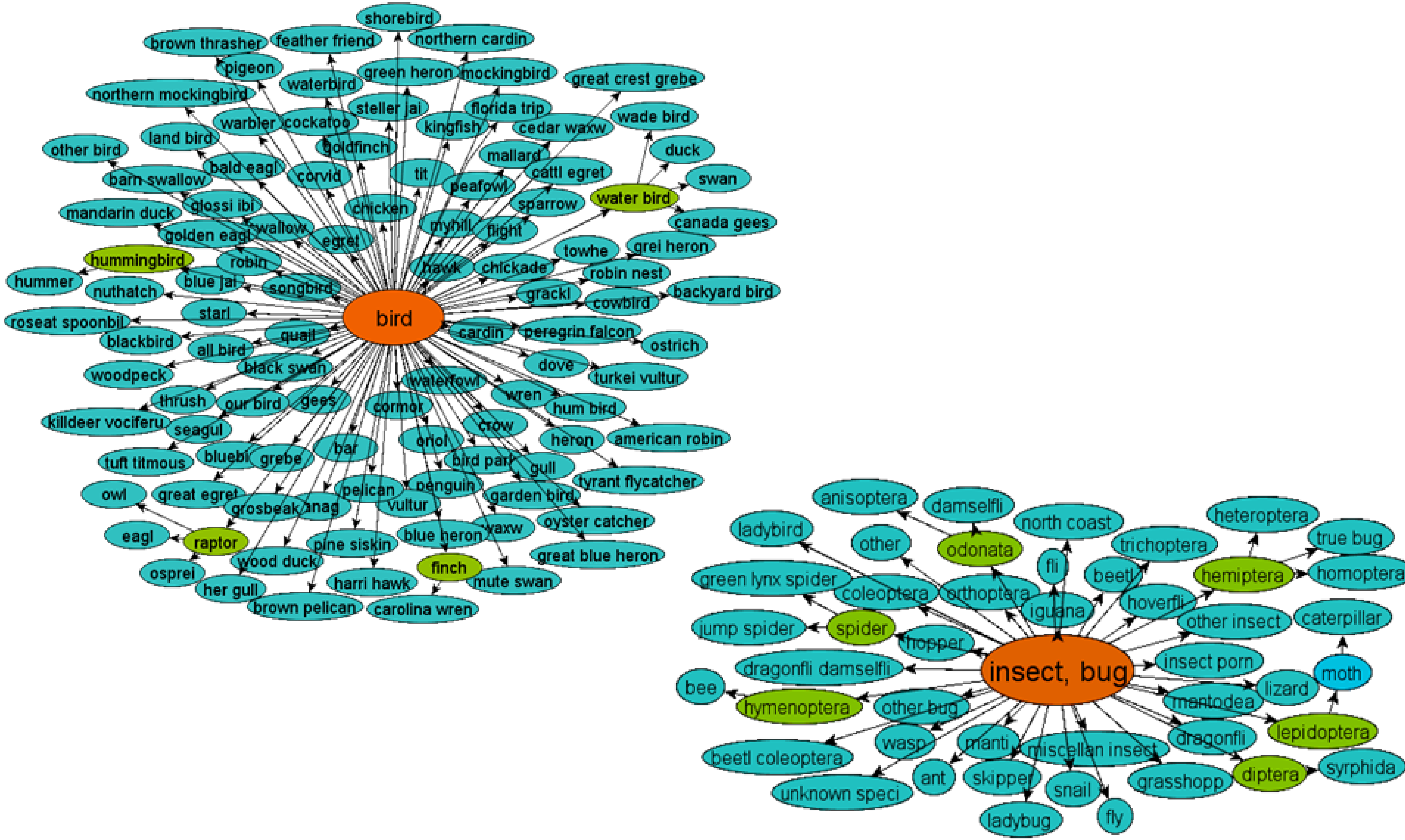
AUT = 5

Experimental Results

Evaluate on 32 cases

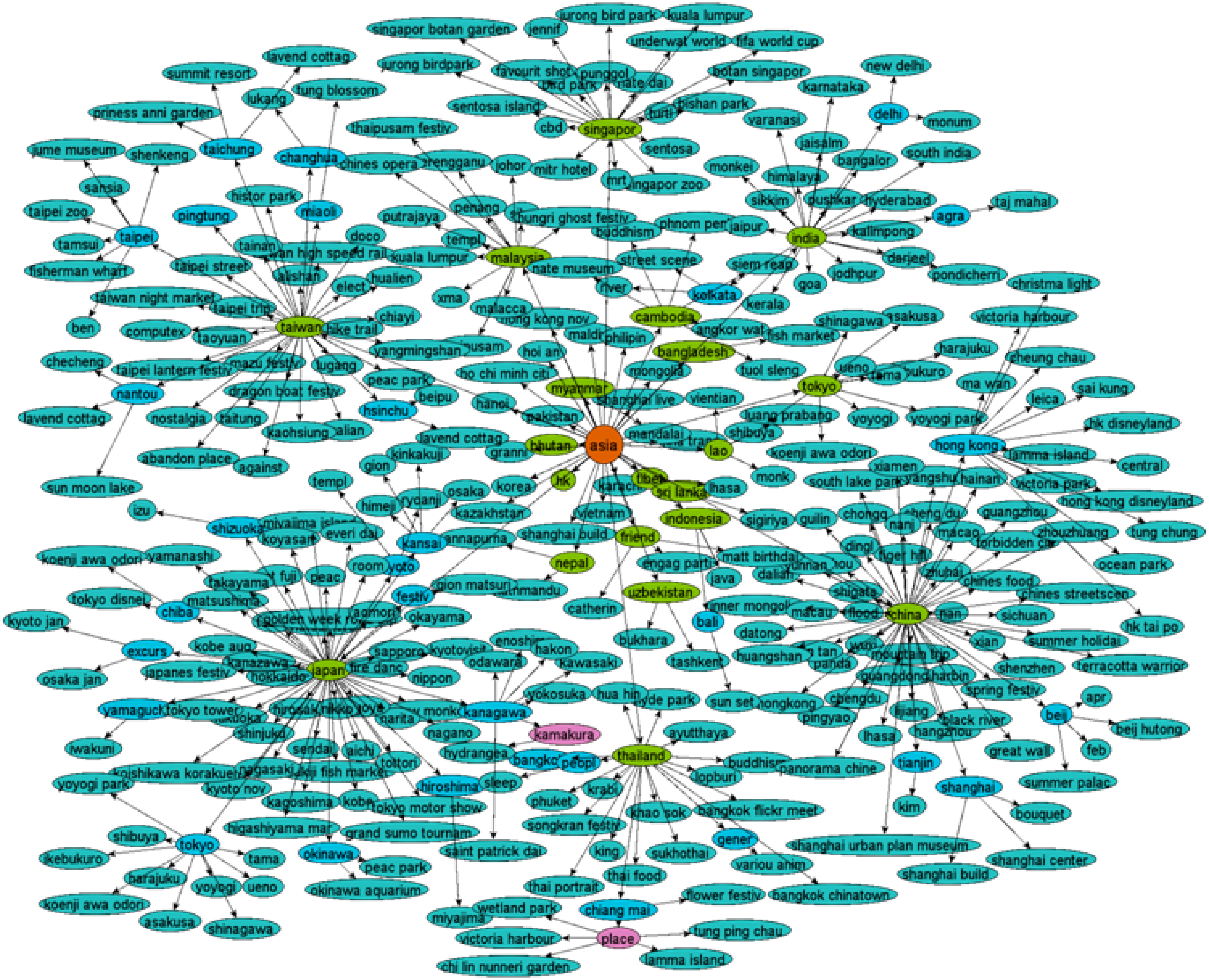
Metrics	# of cases that are superior to the other approach	
	Baseline	The present work
Taxonimic Overlap	7	15
Lexical Recall	6	19
AUT	3	18
TO+LR+AUT	0	11
Accuracy (Manual)	5	5

Some of the Learned Folkonomies



Terms are stemmed

Some of the Learned Folkonomies



Related Work

- Learning concept hierarchy from text data
  - Syntactic based [Hearst92, Caraballo99, Pasca04, Cimiano+05, Snow+06]
  - Word clustering [e.g., Segal+02, Blei+03]
- Induce concept hierarchy from tags
  - Graph-based & clustering based [Mika05, Brooks+06, Heymann+06, Zhou07+]
  - Probabilistic subsumption [Schmitz06]
- Ontology alignment [e.g., Udrea+07]
- Exploit user-specified hierarchy for recommendation
  - GiveALink [Markines06+]

Discussion & Conclusions

- **The present work can create more accurate and more detailed folksonomies than the current state-of-the-art approach, since it exploits structural information during the merging process**

- **The present work is more scalable: incrementally growing the folksonomies rather than using on an exhaustive search**

- **Future work:**
  - ***Automatically separate broader/narrower from related-to relations (facets)***
  - ***combining more sources of evidence such as geographical information***
  - ***Apply on different data sets: e.g., personal workspaces, semantic network***