Correlation Clustering & Crowdsourcing

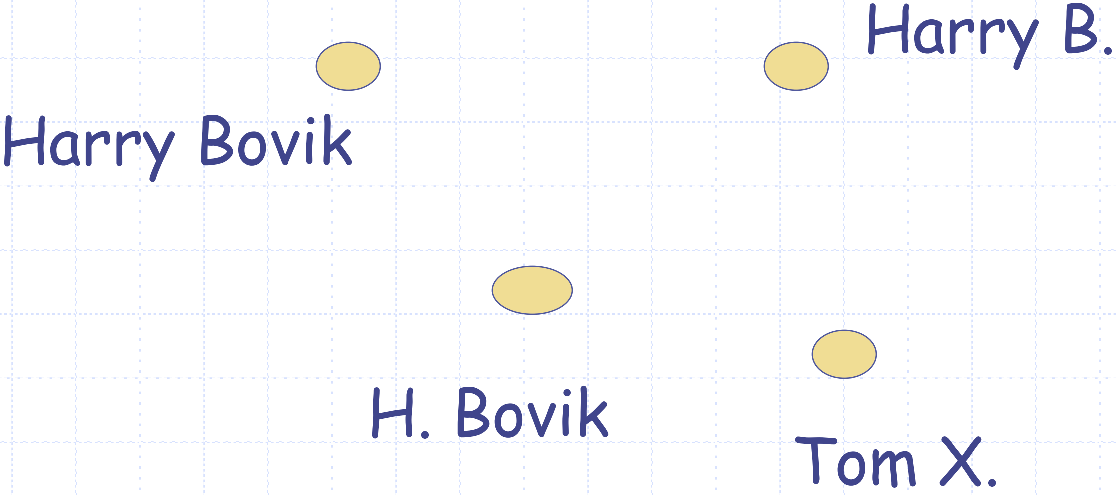
Barna Saha

Acknowledgement: Some slides taken from a presentation by Nikhil Bansal

Clustering

- Say we want to cluster *n* objects of some kind (documents, images, text strings)
- But we don't have a meaningful way to project into Euclidean space.
- Using past data train up some classifier *f(x,y)=same/different*.
- Then run *f* on all pairs and try to find most consistent clustering.

The problem



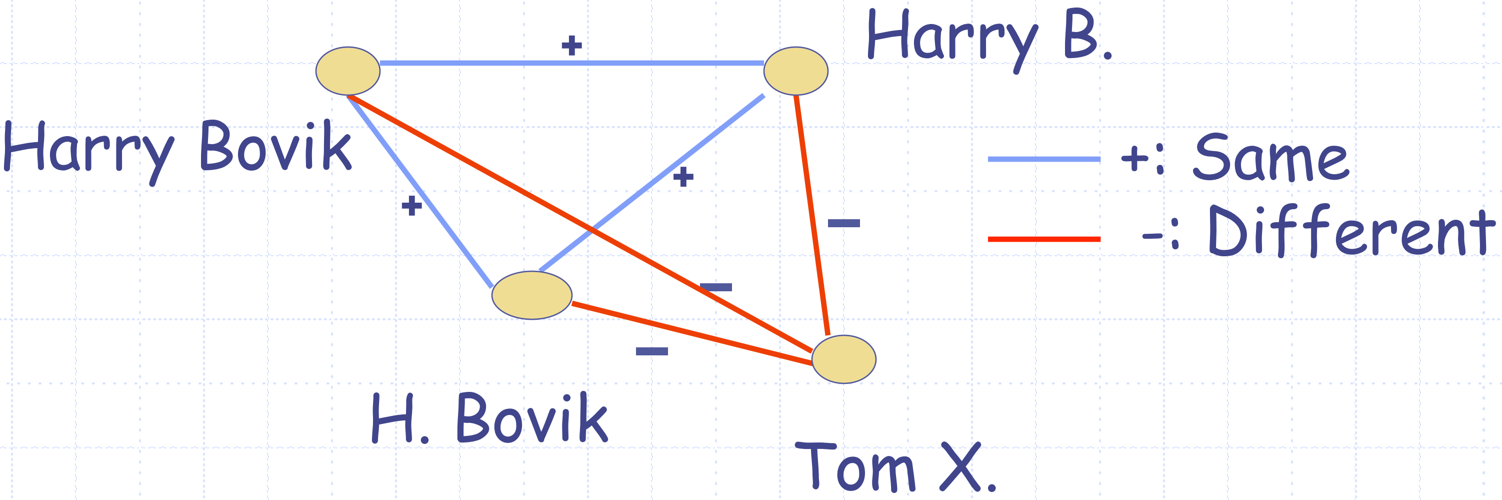
Harry B.

Harry Bovik

H. Bovik

Tom X.

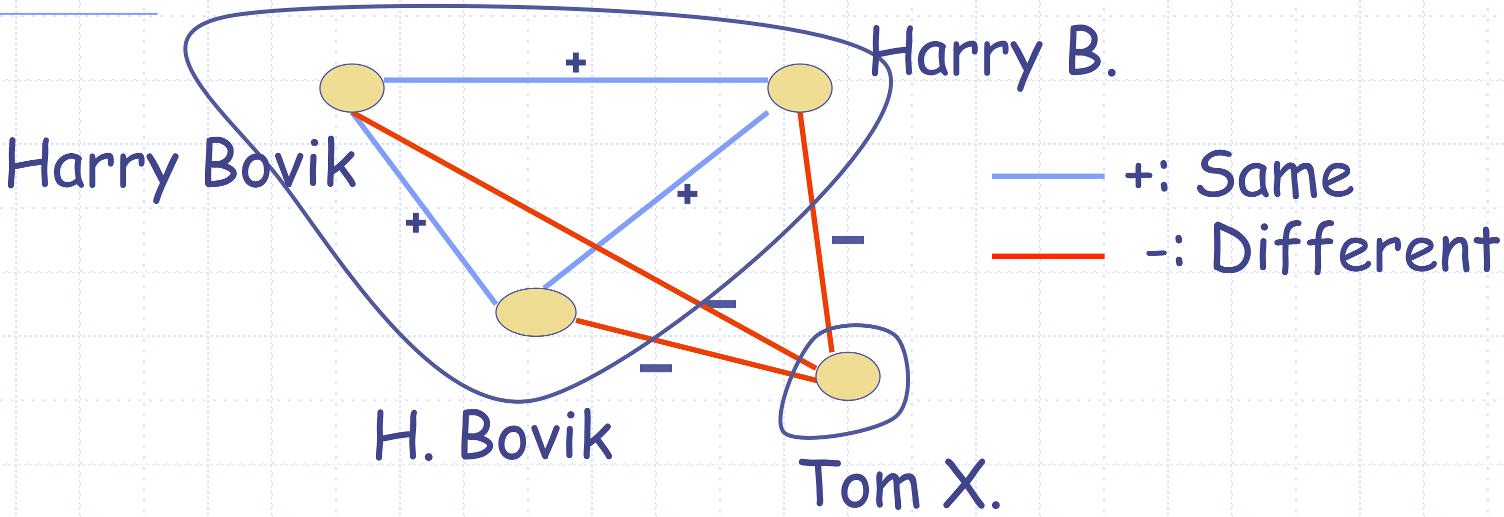
The problem



Train up f(x)= same/different

Run f on all pairs

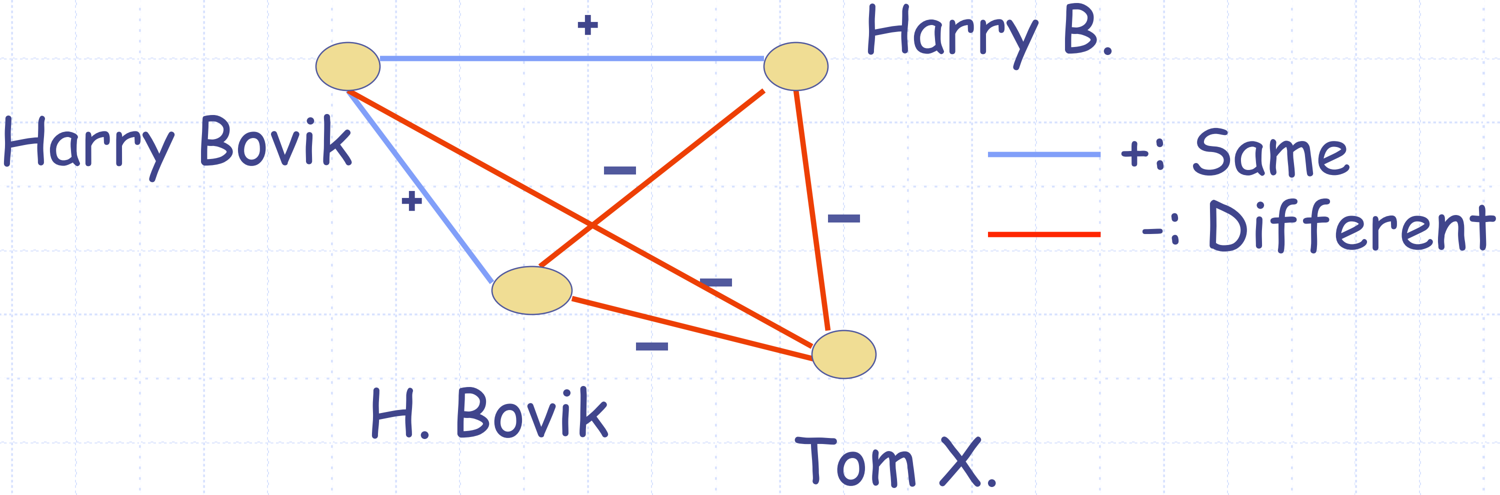
The problem



Totally consistent:

1. + edges inside clusters
2. - edges outside clusters

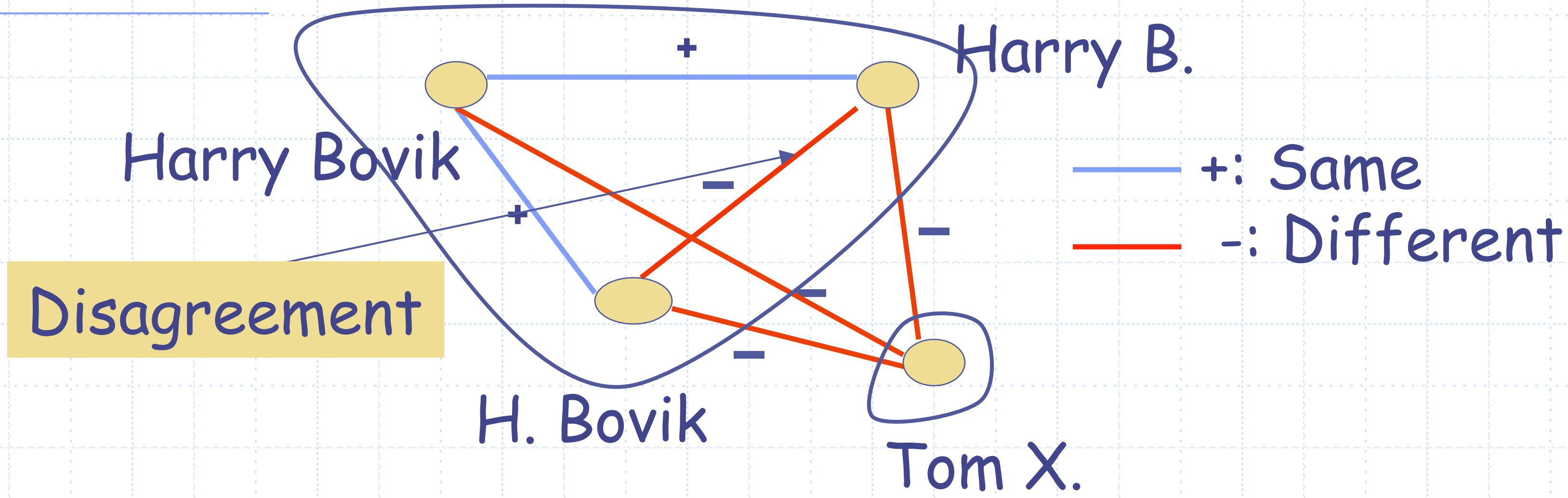
The problem



Train up f(x)= same/different

Run f on all pairs

The problem

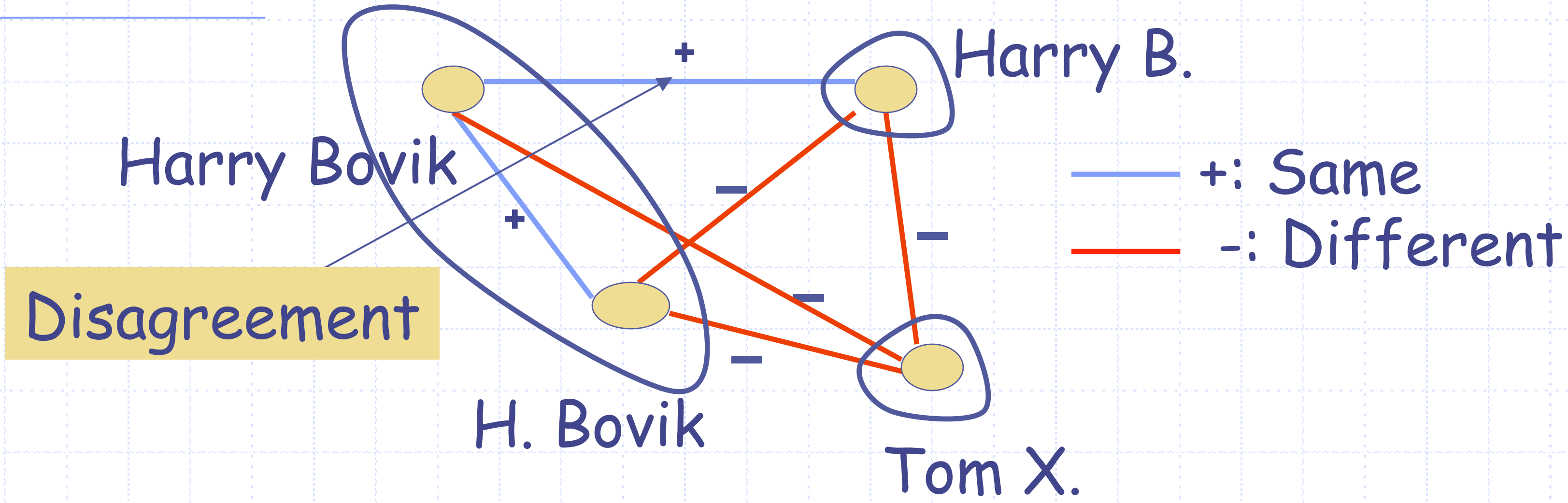


Train up f(x)= same/different

Run f on all pairs

Find most consistent clustering

The problem

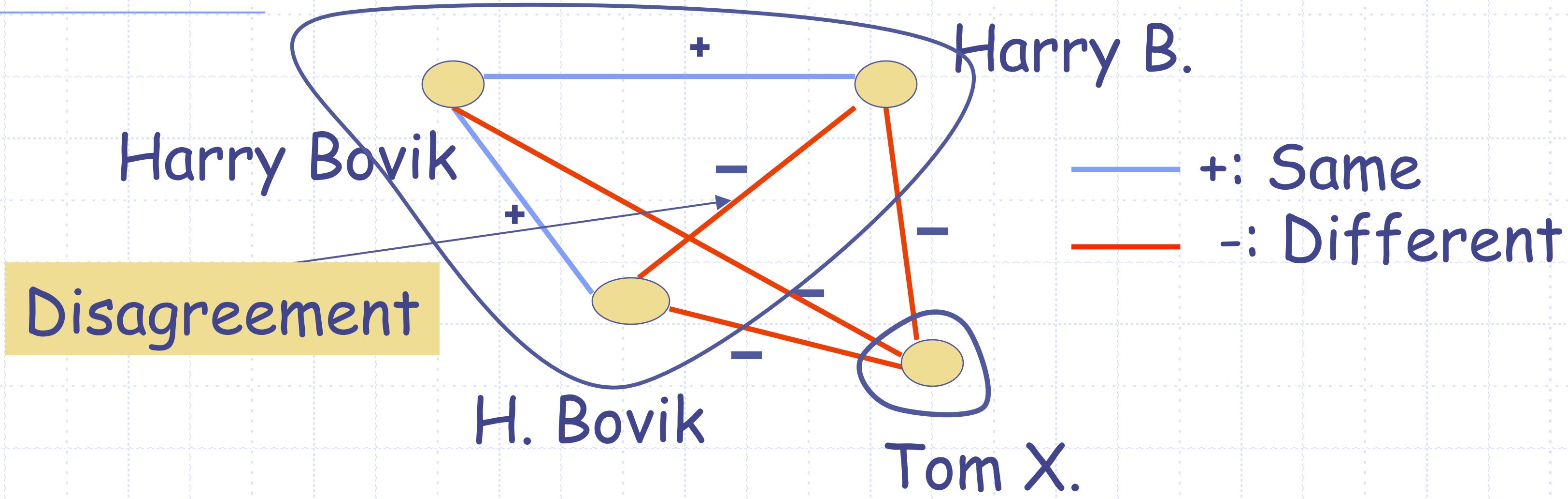


Train up f(x)= same/different

Run f on all pairs

Find most consistent clustering

The problem



Problem: Given a complete graph on n vertices.

Each edge labeled + or -.

Goal is to find partition of vertices as consistent as possible with edge labels.

Max #(agreements) or Min #(disagreements)

There is no k : # of clusters could be anything

The Problem

Noise Removal:

There is some true clustering. However some edges incorrect. Still want to do well.

Agnostic Learning:

No inherent clustering.

Try to find the best representation using hypothesis

Eg: Research communities via collaboration graph

Nice features of formulation

- There's no *k*. (OPT can have anywhere from 1 to n clusters)

- If a perfect solution exists, then it's easy to find: $C(v) = N^{+}(v)$.

- Easy to get agreement on $\frac{1}{2}$ of edges.

PTAS for maximizing agreements

Easy to get ½ of the edges

Goal: additive apx of εn^2 .

Standard approach:

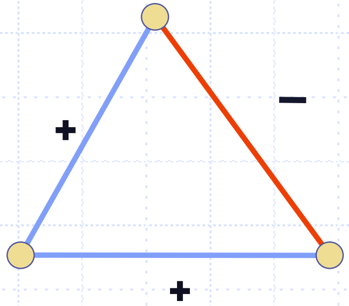
- Draw *small* sample,
- *Guess* partition of sample,
- *Compute* partition of remainder.

Minimizing Disagreements

Goal: Get a constant factor approx.

Lower bounding idea: bad triangles

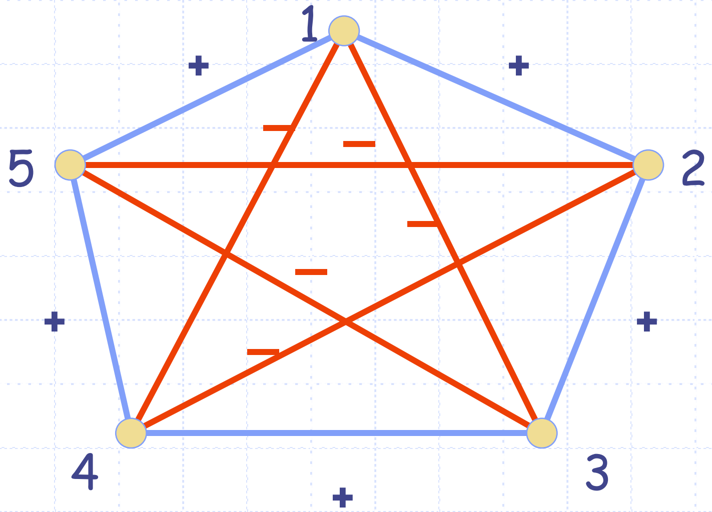
Consider



We know any clustering has to disagree with at least one of these edges.

Lower bounding idea: bad triangles

If several such disjoint, then mistake on each one

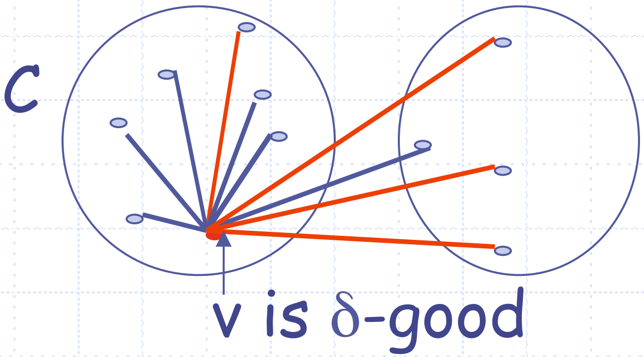


2 Edge disjoint
Bad Triangles
(1,2,3), (3,4,5)

D_{opt} >= #{Edge disjoint bad triangles}

δ-clean Clusters

Given a clustering, vertex δ-good if few disagreements



$N^{-}(v)$ Within C ‹ $\delta|C|$

$N^{+}(v)$ Outside C ‹ $\delta|C|$

— +: Similar

— -: Dissimilar

Essentially, $N^{+}(v)$ ¼ $C(v)$

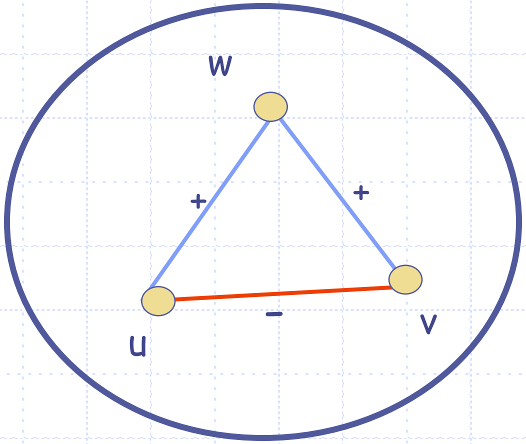
Cluster C δ-clean if all v2C are δ-good

Observation

Any δ-clean clustering is 8 approx for δ<1/4

Idea: Charging mistakes to bad triangles

Intuitively, enough choices of w for each wrong edge (u,v)

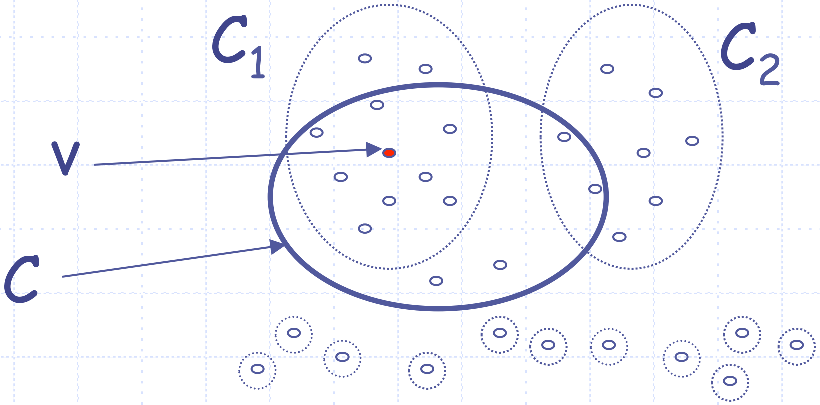


Algorithm

1. Pick vertex v. Let C(v) = $N^{+}(v)$
2. Modify C(v)
   (a) Remove 3δ-bad vertices from C(v).
   (b) Add 7δ good vertices into C(v).
3. Delete C(v). Repeat until done, or above always makes empty clusters.
4. Output nodes left as singletons.

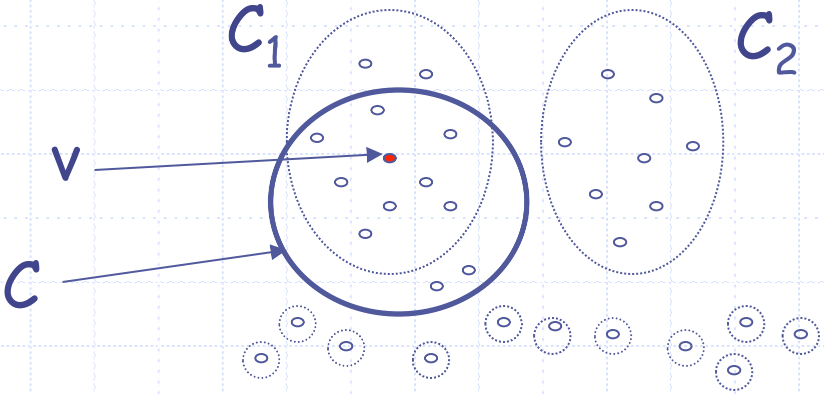
Step 1

Choose v, C= + neighbors of v



Step 2

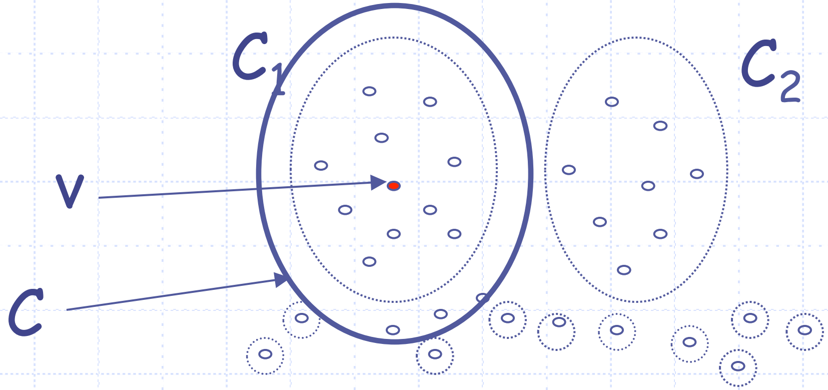
Vertex Removal Phase: If x is 3δ bad, $C=C-\{x\}$



1) No vertex in C_1 removed.
2) All vertices in C_2 removed

Step 3

Vertex Addition Phase: Add 7δ-good vertices to C



1) All remaining vertices in C_1 will be added
2) None in C_2 added
3) Cluster C is 11δ-clean

Extensions

- Better approximation factors very close to 2 are known for minimizing disagreement
- Extensions to weighted case, and arbitrary graph are possible
- For complete graph but with weights O(1)-approximation factor is known
- For weighted arbitrary graphs, O(logn) is the best factor

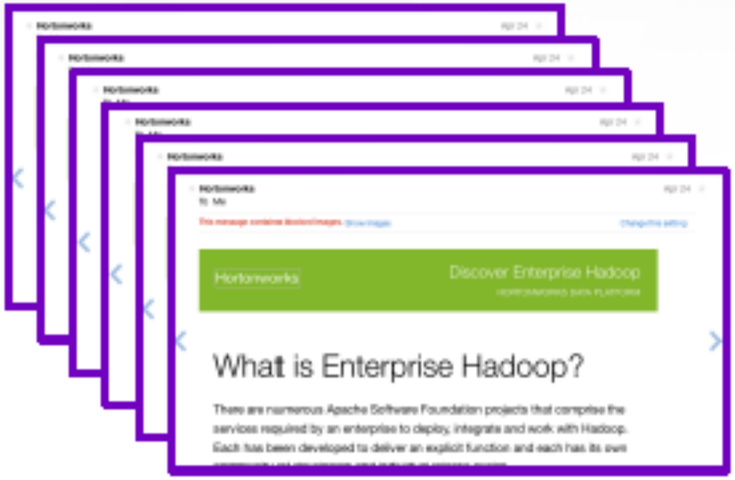
Crowdsourcing

- Using human intelligence to do difficult/ambiguous task
- Many crowdsourcing platform: e.g. Mechanical Turk
- Question: How can we use human intelligence in an effective way?
  - Less cost
  - Better result

Clustering with the crowd

- Unknown disjoint clusters
- Also known as Entity resolution/ Deduplication/Record linkage etc.

Document de-duplication



Hortonworks

Hortonworks

Hortonworks

Hortonworks

Hortonworks

Hortonworks

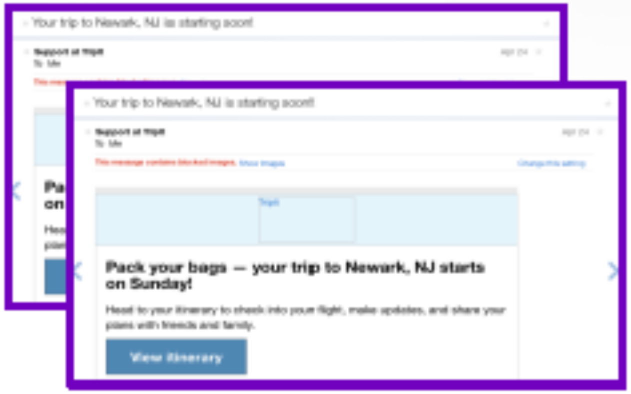
To: Me

Hortonworks

Discover Enterprise Hadoop

What is Enterprise Hadoop?

There are numerous Apache Software Foundation projects that comprise the services required by an enterprise to deploy, integrate and work with Hadoop. Each has been developed to deliver an explicit function and each has its own



Your trip to Newark, NJ is starting soon!

Your trip to Newark, NJ is starting soon!

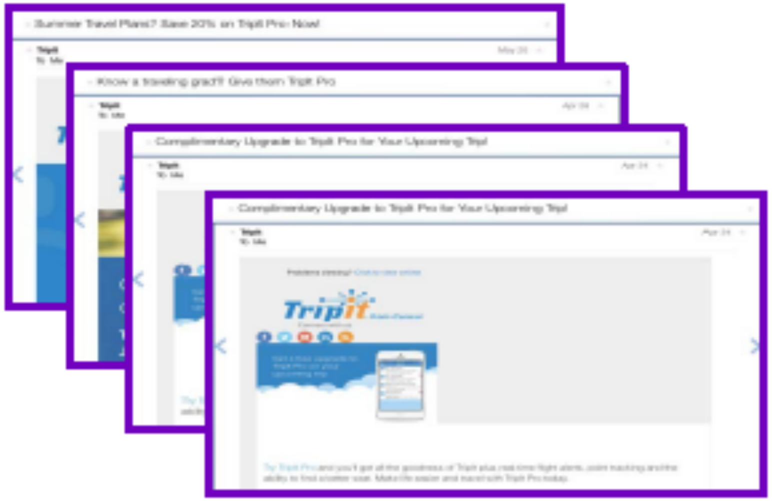
Support at TripIt

To: Me

Pack your bags — your trip to Newark, NJ starts on Sunday!

Head to your itinerary to check into your flight, make updates, and share your plans with friends and family.

View itinerary



Summer Travel Plans? Save 20% on TripIt Pro Now!

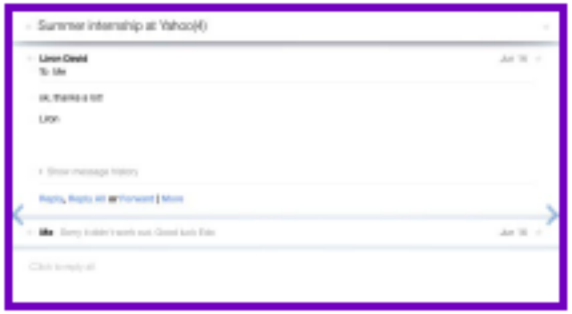
Know a traveling grad? Give them TripIt Pro

Complimentary Upgrade to TripIt Pro for Your Upcoming Trip!

Complimentary Upgrade to TripIt Pro for Your Upcoming Trip!

TripIt

To: Me



Summer internship at Yahoo(4)

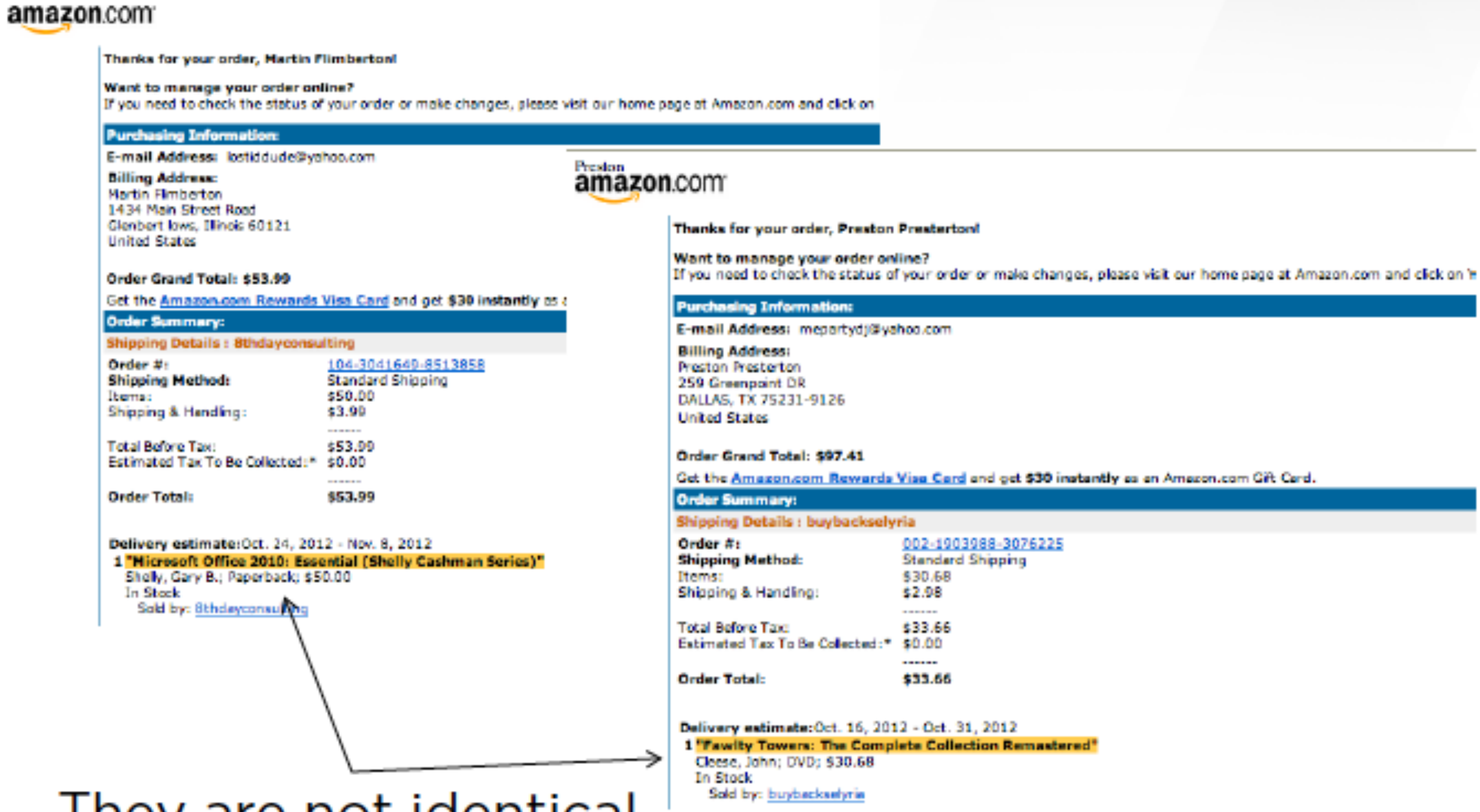
To: Me

Show message history

Reply, Reply All or Forward | More

Click to reply all

Document de-duplication



amazon.com

Thanks for your order, Martin Flimberton!

Want to manage your order online?
If you need to check the status of your order or make changes, please visit our home page at Amazon.com and click on

Purchasing Information:

E-mail Address: lostiddude@yahoo.com

Billing Address:
Martin Flimberton
1434 Main Street Road
Glenbert lows, Illinois 60121
United States

Order Grand Total: $53.99

Get the Amazon.com Rewards Visa Card and get **$30 instantly** as

Order Summary:

Shipping Details : 8thdayconsulting

Order #:	104-3041640-8513858
Shipping Method:	Standard Shipping
Items:	$50.00
Shipping & Handling:	$3.99

Total Before Tax:	$53.99
Estimated Tax To Be Collected:*	$0.00

Order Total:	**$53.99**

Delivery estimate: Oct. 24, 2012 - Nov. 8, 2012
1 "Microsoft Office 2010: Essential (Shelly Cashman Series)"
Shelly, Gary B.; Paperback; $50.00
In Stock
Sold by: 8thdayconsulting

Preston
amazon.com

Thanks for your order, Preston Presterton!

Want to manage your order online?
If you need to check the status of your order or make changes, please visit our home page at Amazon.com and click on

Purchasing Information:

E-mail Address: mepartydj@yahoo.com

Billing Address:
Preston Presterton
259 Greenpoint DR
DALLAS, TX 75231-9126
United States

Order Grand Total: $97.41

Get the Amazon.com Rewards Visa Card and get **$30 instantly** as an Amazon.com Gift Card.

Order Summary:

Shipping Details : buybackselyria

Order #:	002-1903988-3076225
Shipping Method:	Standard Shipping
Items:	$30.68
Shipping & Handling:	$2.98

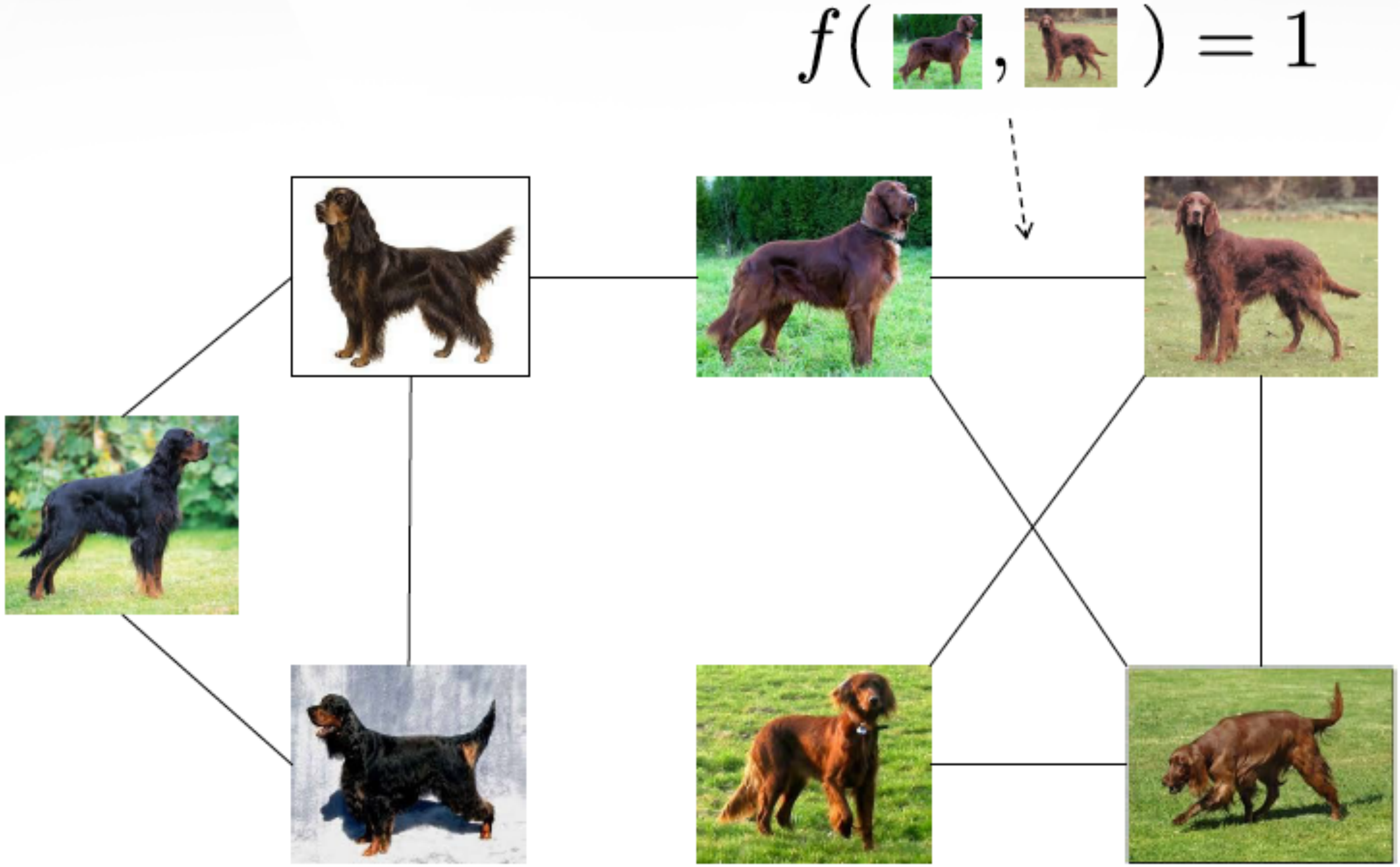
Total Before Tax:	$33.66
Estimated Tax To Be Collected:*	$0.00

Order Total:	**$33.66**

Delivery estimate: Oct. 16, 2012 - Oct. 31, 2012
1 "Fawlty Towers: The Complete Collection Remastered"
Cleese, John; DVD; $30.68
In Stock
Sold by: buybackselyria

They are not identical

Motivation from machine learning



Minimizing number of queries

- We have some prior believes f(i,j) in [0,1]
- Closer to 1 means i and j are likely to belong to the same cluster
- Closer to 0 means i and j are likely to belong to different clusters
- Design a strategy to ask minimum number of pair-wise queries to the crowd to recover the true clustering