

# Information and Beyond

Becky Robilotto

Group Meeting

March 22, 2011

# Outline

- Content Management Systems (CMS)
- Culturomics – Investigating Twitter

# CMS


## Content Management System


# CMS info from wikipedia


- Allow for a large number of people to contribute to and share stored data
- Control access to data, based on user roles (defining which information users or user groups can view, edit, publish, etc.)
- Aid in easy storage and retrieval of data
- Reduce repetitive duplicate input
- Improve communication between users

# Different Examples

- Drupal
- Joomla
- Wordpress
- What one is the best?
  - Is there a way to connect all different forms of communication
    - etherpad + wiki + email + google docs + twitter.

-  [WORDPRESS.ORG](https://WordPress.org) - Best for blogging, easy to set up and has easily used widgets and themes. Over 202 million websites use it.

-  – Useful in the backend of things, like registering users, more for the developer. Confusing for a new programmer. Harder to change the design. Not as easy to blog. Users can be tracked. MTV, BBC, The Onion uses Drupal.

-  [Joomla!](https://Joomla.org) – More designer friendly and easy to customize for the themes. It is not Web 2.0 compatible. Lacking in social and user oriented areas.

- Drupal > Joomla > Wordpress

# Developing Pseudogene.org 2.0

- Utilizing Drupal, create a new pseudogene resource that is easily updated and can incorporate more content
  - Possible blog posts describing updates
  - Easy ways to link to other resources, such as WormBase.org, FlyBase.org, Ensembl
  - Put in a twitter feed, with hashtag #pseudogene
  - Allow for user IDs for editing

## Drupal Demo

Was planning to show backbone, but I  
accidentally screwed up configuration  
and it is not working,  
so I will show a known example...



# Future Goals

- Find possible new interesting modules
- Develop new modules that are more specific to pseudogenes
- Develop a better pseudogene.org

# Culturomics

- Culturomics: the application of high-throughput data collection and analysis to the study of human culture.
- Using the Google n-gram, investigated 4% of all books printed
- Looked for insights in lexicography, evolution of grammar, and adoption of technology, etc.
- Could a similar study be done using the realtime data from twitter?

(Science 2010)

# Time Scale of Information Disbursement



Social Networking Sites

Blogs

Twitter  
Facebook  
Scienceblogs.com

Articles

Magazines,  
Newspapers

Books

With Nitin and Koon-Kiu

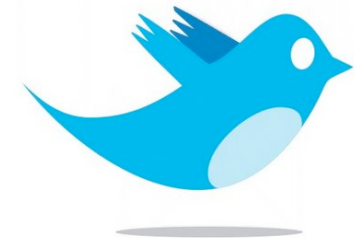
# twitter

- Celebrated its 5<sup>th</sup> birthday March 21
- Twitter is a **real-time**, fluid information network
  - 175 million registered users
  - 140M tweets are written per day.
- Each Tweet is 140 characters in length
  - Headlines, photos, articles, video
- Different ways people are tweeting
  - News, Spam, Self-promotion, Pass-along value, Conversational, Pointless babble
- Some people just get value from following others



# How did I get interested?

- Originally thought it was for mindless celebrity stalking
  - Top followed: Lady Gaga, Justin Bieber, Britney Spears, Barack Obama, Kim Kardashian
- Discovered how useful it was when I went to the Beyond the PDF workshop
- Utilized hashtags #beyondthepdf so people could follow the discussion
- There was legitimate scientific discussion occurring freely and openly
- During group discussion, they had the “twitter waterfall” on the projector screen
- Allows for short concise ideas, one at a time
- How could all this information be utilized?





# Three Different Approaches

- Build a network based off of recognized twitter accounts
- Build a network based off of certain keyword terms
- Utilize Resource Description Framework (RDF) to create a metadata data model based on keyword terms

# Parsing tweets

- Hashtags – can look up interesting topics of interest #beyondthepdf
- @replies – conversations between people
- Retweets RT – a interesting tweet that has been shared by other people
- Meta-data – location, description, time analysis



[ElsevierLabs](#) Thanks [@pgroth](#) helped me find the 'Searches' tab for [#BeyondthePDF!](#)

Twitter - Jan 19, 2011 12:30:55 PM

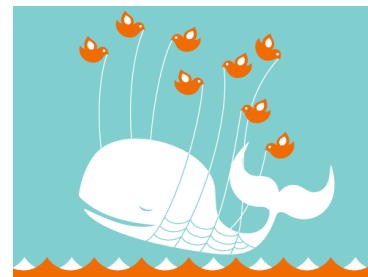


[GigaScience](#) Here is the paper ... RT [@petemurrayrust](#): [#beyondthepdf](#) Heather Piwowar showed that publishing data gets more citations

[PLoS ONE: Sharing Detailed Research Data Is Associated with ...](#) - plosone.org

# Archiving the tweets

- This is a problem because the tweets seem to be searchable for about a month or less
- Applications available for archiving the tweets
  - <http://twapperkeeper.com/>
  - <http://archivist.visitmix.com/>
  - <http://searchtastic.com/export.php>
  - <http://www.google.com/realtime>
- **Twitter changed its terms and services of its API, and it doesn't seem to allow for archiving of individual tweets anymore**





# Google Realtime

- Realtime Search lets you see up-to-the-second social updates, news articles and blog posts about hot topics around the world
  - Will search twitter archives
- Displays full conversations
- Also displays timeline and you can pinpoint the particular tweet of interest



# Archivist – will still do analytics

## Archive: #beyondthepdf

Remove archive ✕

Archive contains  
**1,152 Tweets**

Archive started  
**1/20/2011**

Archive status  
**Archiving**

Archive last updated  
**8 hours ago**

Visualizations updated  
**13 hours ago**

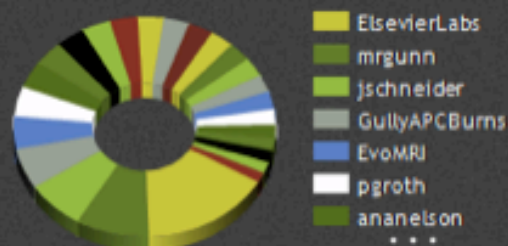
Archive is  
**Private** 🔒

\*\*\*The data displayed may not include all tweets with the search term. Read the [FAQ](#) to understand more.\*\*\*

### Tweet Volume Over Time



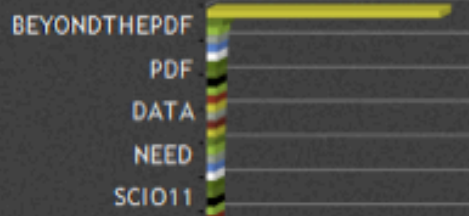
### Top Users



### Tweet Vs. ReTweet



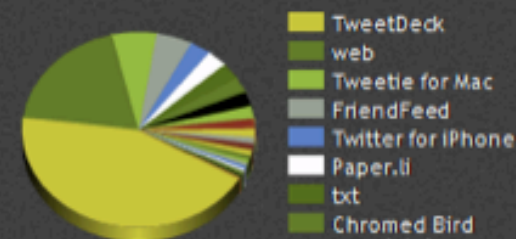
### Top Words



### Top Urls





### Source

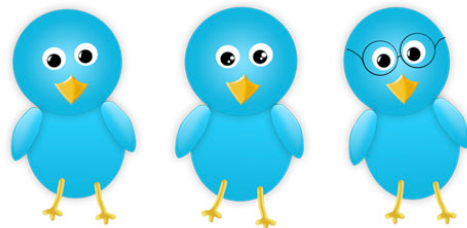


# Problems of searching Twitter

- Different languages – may not get relevant tweets
- Use of non-standard abbreviations
- Links may use shortened URLs, one may not know what it goes to
- Tags may not have the same meaning
  - Example #CBB (Computational Biology and Bioinformatics) or (Can't Be Bothered)

# Building a network based on users

- Chose two well know usernames
- @sciencemagazine - Science Magazine 
  - 15,674 followers
- @NatureNews– Nature News 
  - 143,317 followers
- Can we see how ideas/terms are propagated through the network?







# Building a network on keyword terms

- Picked a biological term to start with
- Picked the term pseudogene and created a word cloud based on frequency
- Parse out keyword terms
  - Hard – need to think of optimum way
- Build a matrix based on co-occurrence





# Adjacency Matrix

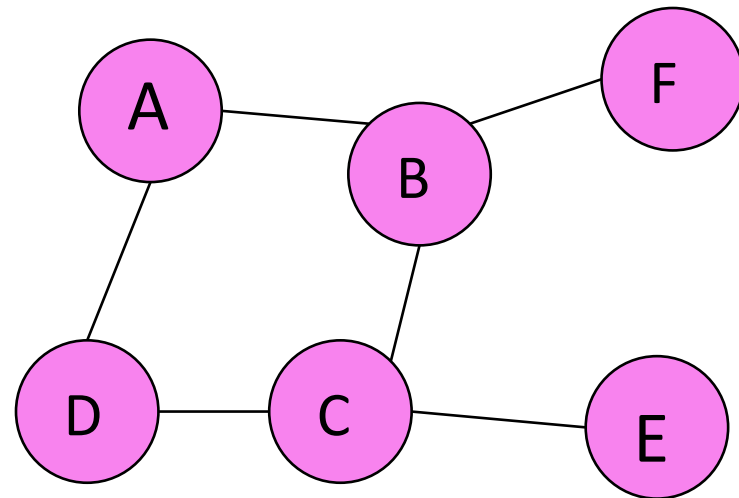
It symbolizes a graph of  $n$  nodes or vertices

Each row and each column corresponds to different vertices

$X_{i,j}$  represents a connection in row  $i$  and column  $j$

If two elements are connected a 1 will appear in  $X_{i,j}$  or else a 0 will appear

	A	B	C	D	E	F
A	0	1	0	1	0	0
B	1	0	1	0	0	1
C	0	1	0	1	1	0
D	1	0	1	0	0	0
E	0	0	1	0	0	0
F	0	1	0	0	0	0



*Carrano, Savitch Data Structures and Abstractions with Java(2003)*

# Simple graph approach

- Create adjacency matrix based on terms mentions in the tweet
  - May not be the most efficient approach
  - Examples
    - pseudogene (P) and function (F)
    - pseudogene (P) and codon (C)
    - codon (C) and intron (I)
    - pseudogene (P) and intron (I)
- Want to build a circular connection

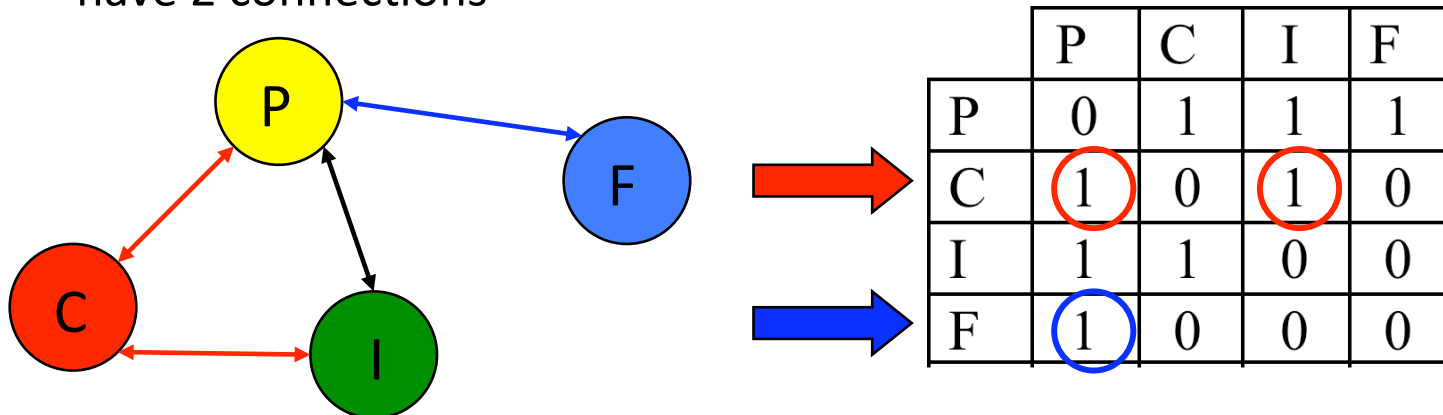
# Solutions

Couldn't use different traversal algorithm because it would have same result of printing out everything.

Focused the adjacency matrix to see if adjustments could be made

Noticed that the when the nodes are in the connection they are connected to 2 or more other nodes and when the node was just attached, it was connected to 1 node.

Idea is to go through the adjacency matrix and get rid of all the vertices that only have 1 connection and continue until all the vertices have 2 connections

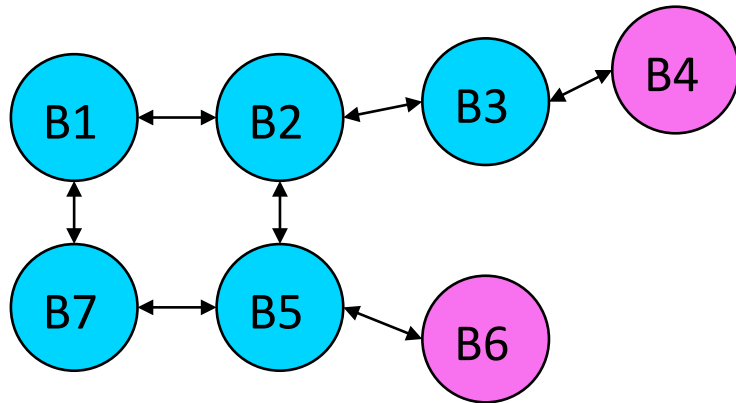


# Tracing the algorithm

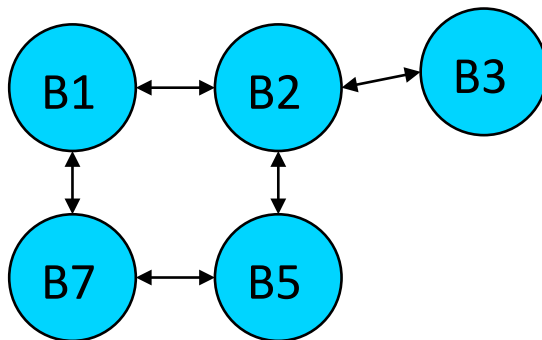
Nodes with two or more 1s = 5

Total nodes = 7

Delete nodes with one 1



	B1	B2	B3	B4	B5	B6	B7
B1	0	1	0	0	0	0	1
B2	1	0	1	0	1	0	0
B3	0	1	0	1	0	0	0
B4	0	0	1	0	0	0	0
B5	0	1	0	0	0	0	1
B6	0	0	0	0	1	0	0
B7	1	0	0	0	1	0	0



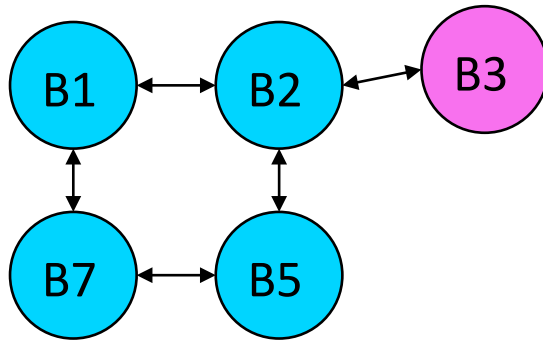
	B1	B2	B3	B5	B7
B1	0	1	0	0	1
B2	1	0	1	1	0
B3	0	1	0	0	0
B5	0	1	0	0	1
B7	1	0	0	1	0

# Tracing the algorithm cont.

Nodes with two or more 1s = 4

Total nodes = 5

Delete nodes with one 1

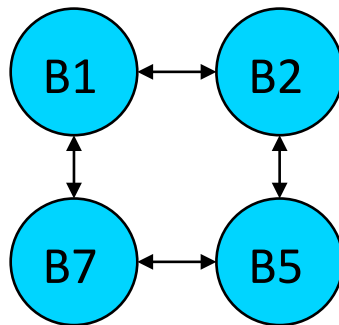


	B1	B2	B3	B5	B7
B1	0	1	0	0	1
B2	1	0	1	1	0
B3	0	1	0	0	0
B5	0	1	0	0	1
B7	1	0	0	1	0

Nodes with two or more 1s = 4

Total nodes = 4

Done. Run Depth-First Traversal



	B1	B2	B5	B7
B1	0	1	0	1
B2	1	0	1	0
B5	0	1	0	1
B7	1	0	1	0

Prints out: B1 B2 B5 B7

# Using this algorithm

- Utilizes this type of algorithm, we may be able to infer some new information
- Our example would show pseudogene, intron, and codon occur together
- Problems:
  - We may not be able to parse the keywords well enough
  - There may not be enough co-occurrences to build good connections

# Resource Description Framework Triples

- Each RDF triple is made up of subject, predicate, and object.
- Each RDF triple is a complete and unique fact.
- Each RDF triple can be joined with other RDF triples, but it still retains its own unique meaning, regardless of the complexity of the models in which it is included.

# Parse into RDF



[mahler83](#) RT [@science3point0](#): Why dont all **journals** have a **data link** for the raw **data?** [#beyondthepdf](#)  
Twitter - Jan 19, 2011 12:02:30 PM



[mfenner](#) Paul Groth: **provenance** essential **part** of scientific **process**. How do we integrate into paper or future? [#beyondthepdf](#)  
Twitter - Jan 19, 2011 12:05:44 PM



[laurajcroft](#) how do we avoid annotated **documents** becoming a 'christmas tree of **hyperlinks**'? good question [#beyondthepdf](#)  
Twitter - Jan 19, 2011 12:51:55 PM



Subject



Predicate



Object

Use this to build an Entity Relationships, or ontologies



# Impossible Task

- Can we make the popularity of Twitter more useful?
- Nanopublishing?
- Could we get people to tweet in a more structured way, for easier parsing?
  - Impossible, because it is already limited to 140 characters
  - Not likely to change style of tweeting

# Future Goals

- Find a good way to archive tweets and hashtags
- Find a better way to parse the tweets
- Can we expand similar analysis to blogs as well as tweets?
- Can expand keywords to more than one biological term?
- Search locations, times, number of retweets for more layers and look for dynamics
- Find a way to link tweets to websites

# Acknowledgments

- Koon-Kiu
- Nitin
- Genome Annotation
- Genome Tech
- Kei Cheung
- Mark Gerstein

# Questions?

