Dataverse Network (DVN)
Patent Network Database Project

Harvard Business School
Institute for Quantitative Social Science, Harvard University

University of Trento – Thursday, July 16, 2009

Ronald Lai rolai@hbs.edu
Alex D'Amour adamour@iq.harvard.edu
Lee Fleming lfleming@hbs.edu

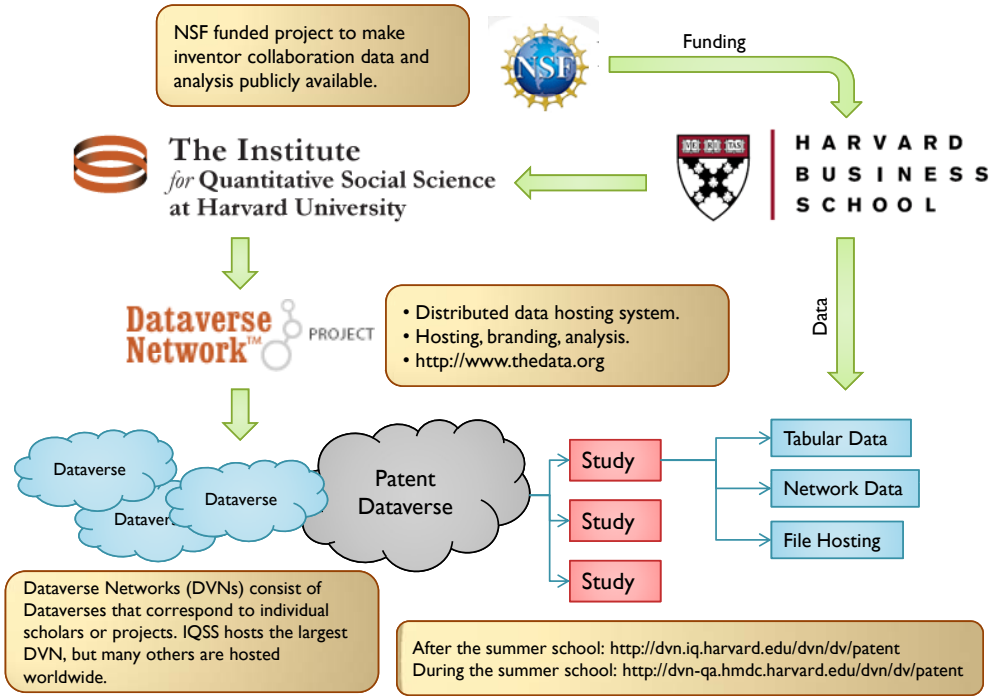
Access Our Dataverse at: http://dvn.iq.harvard.edu/dvn/dv/patent

We would like to thank the NSF for supporting this research. Errors and omissions remain ours (though we ask that you bring them to our attention).

DVN Patent Network Database Project

INTRO DATAVERSE NETWORK

Institutional Support



Roadmap

- **PRELUDE** PATENT DATA
  - Patent data. (Flat data, network data)
  - Inventor disambiguation.
  - ☆ *Brief* overview, hold questions until lab time.

- **FROM** RESEARCH INTEREST **TO** REGRESSIONS
  - Patents – Why? How? Now what?
  - ☆ Lab time to test drive DVN network tools.

DVN Patent Network Database Project

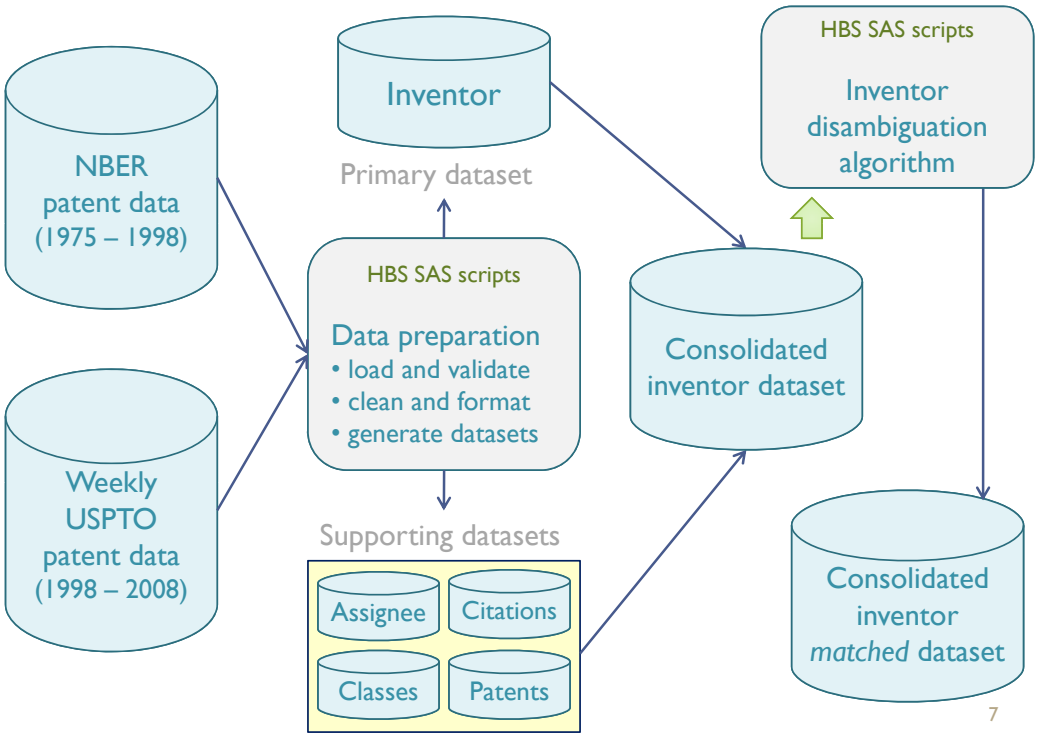
PRELUDE PATENT DATA

Patent Data, A glimpse

Invnum_N	Name	Patent	Assignee	City	Loc	…
12345	Fleming, Lee	5029133	HP	Fremont	CA	…
12345	Fleming, Lee	9999999	Harvard	Cambridge	MA	…
45678	D'Amour, Alex	9999999	Harvard	Boston	MA	…
67890	Lai, Ronald	9999999	Harvard	Randolph	MA	…

- Prof Fleming, Alex, and Ron collaborate on patent 9999999.
  - Data are organized in unique inventor-patent pairs.
    - *Unique inventor number* (HBS disambiguation algorithm), constant between patents.
    - Patent is assigned to one entity (usually inventors' employer or self if blank), constant over a patent.
    - City and Loc are personal addresses (at the city level) of inventors, vary over a patent.

Patent Data, Our process



Patent Data, Disambiguation

- Current method (2008)
  - Linear, unsupervised.
  - Records are compared element by element.
  - Similarity between records is a weighted average of element-wise similarity scores.
  - ▬ Currently, weights are not optimized.
- Future methods (est. 2010)
  - Semi-supervised with automatically generated training sets.
  - Multi-dimensional similarity profiles.
  - Optimal weighting, non linear interactions.
  - ☆ Looking for collaborators for this initiative.

Patent Data, Known issues

- US inventors are better disambiguated
  - Location information is more specific for US.
  - Large name, assignee variation.
  - Thus, Asian inventors are especially difficult.
- Time inconsistency
  - Assignee information changes as a function of mergers, renaming.
  - Zip codes change over time.
- Name inconsistencies
  - Typographical errors.
  - Nicknames and name changes.

Patent Data, Data sources

- HBS patent dataset (incl. documentation for inventor disambiguation algorithm):
  - http://dvn.iq.harvard.edu/dvn/dv/patent/
- NBER patent data
  - http://www.nber.org/patents/
- USPTO patent data (FTP based):
  - ftp://ftp.uspto.gov/pub/patdata/2009/Grants/
- HBS algorithms will be publically available as we encourage **open community** improvement!

Patent Data, Contents

Inventor-Patent Database

**HBS algorithm generated variables.*

INVENTOR	
Invnum_N	*Unique inventor number.*
Name	Inventor name (Last, first middle initial)
Inv_Seq	Inventor number on patent.

PATENT	
Invpatseq	*Generated patent history sequence for inventor.*
Pat_Type	94% Utility (U), 5% Design (D).
Patent	USPTO assigned patent number.
AppDate	Patent application date.
GDate	Patent grant date.
Claims	Total patent claims.

INVENTOR LOCATION	
City	Inventor's city.
Loc	(US only) State or Country code.
Zipcode	(US only) Zipcode.
Usflg	US = 1 (50 states, D.C. and territories)

ASSIGNEE	
Assignee	Primary firm associated with patent.
Numasg	*Generated assignee number.*

CLASSES	
Classes	Classes separated by "|". Main-Sub. Primary class listed first. 7 maximum.
Scls_cnt	*Count of total classes.*
Scls_1	*Flag. Contains a first occurrence of class.*
Scls_pair_1	*Total of first occurence of class-pairs.*

CITATION (Fieldnames combine the concepts)	
Back	*Backward citations (patents cited to)*
Forw	*Forward citation (patents cited from).*
5	*5 year window.(otherwise, for all time).*
SC	*Assignee self citation.*

Creating the Network

Invnum_N	Name	Patent	Assignee	City	Loc	…
12345	Fleming, Lee	5029133	HP	Fremont	CA	…
12345	Fleming, Lee	9999999	Harvard	Cambridge	MA	…
45678	D'Amour, Alex	9999999	Harvard	Boston	MA	…
67890	Lai, Ronald	9999999	Harvard	Randolph	MA	…

Group by *Unique Inventor Identifier.*

Merge with self on patent number.

VERTICES (INVENTORS)

Invnum_N	Name
12345	Fleming, Lee
45678	D'Amour, Alexander
67890	Lai, Ronald

EDGES (PATENT)

Head	Tail	Patent	…
12345	45678	9999999	…
12345	67890	9999999	…
45678	67890	9999999	…

Note: Some information is lost in constructing the network dataset. Only certain variables from the inventor-patent data are recorded in the network. Also, patents of singleton inventors are not recorded. It is necessary to complement the network dataset with the inventor-patent dataset for a full picture.

Network Data

Vertex-Edge Dataset

VERTICES (INVENTORS)	
Invnum_N	*Inventor number on patent.*
Name	Inventor name (Last, first middle initial)

EDGES (PATENT)	
H	*"Head" inventor – inventor on the edge with the smaller id number.* Loc, City, Zip for each inventor is also provided.
T	*"Tail" inventor – inventor on the edge with larger id number.* Loc, City, Zip for each inventor is also provided.
Numasg	*Generated assignee identifier.*
Pat_Type	94% Utility (U), 5% Design (D).
Patent	USPTO assigned patent number.
AppYear	Patent's application year.

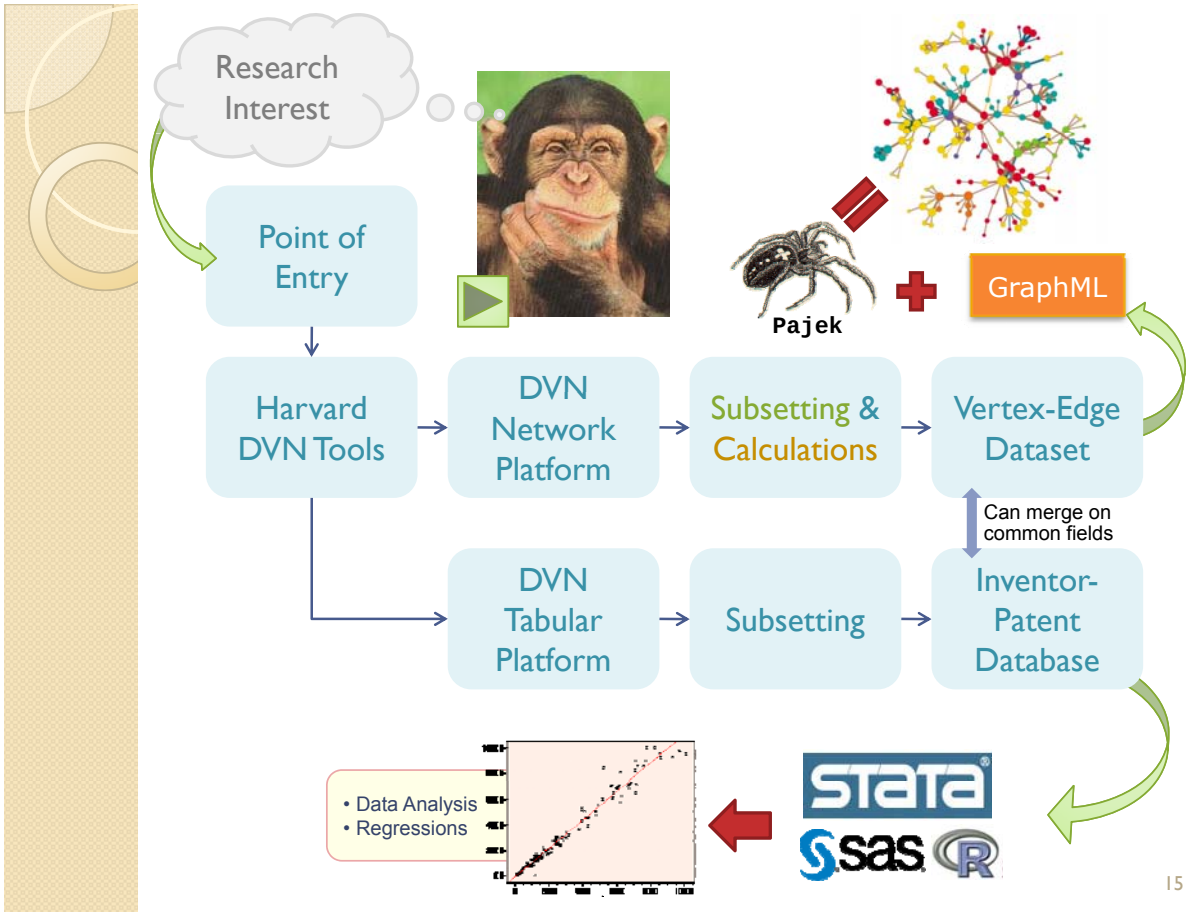
Vertex represent inventors

Note: Due to space and memory constraints, a number of these edge or vertex attributes may be left out from larger networks posted on our Dataverse.

Edges represent patents

DVN Patent Network Database Project

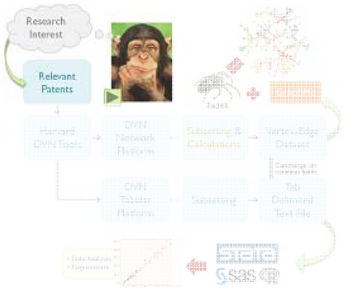
FROM RESEARCH INTEREST TO REGRESSIONS



DVN Patent Network Database Project

STEP I RESEARCH INTEREST

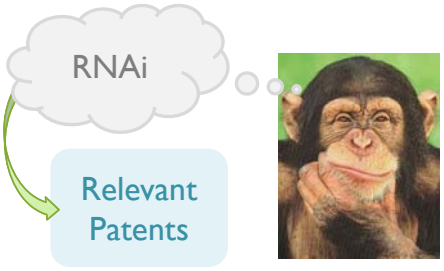
Research Interest



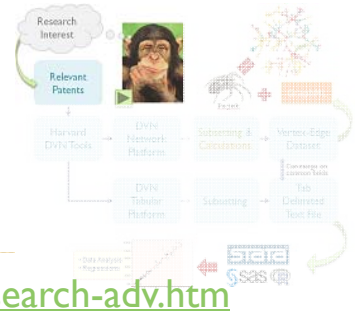
- Applicable Research
  - Origin of breakthroughs
  - Impact of legislation on innovation
  - Organizational influence on innovation

Example: How did breakthroughs in RNAi technology take place?

Challenge: Mapping this idea (RNAi) onto a point of entry for the inventor network.



Research Interest

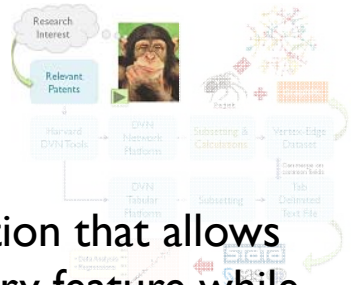


- Patents as a point of entry
- http://patft.uspto.gov/netahtml/PTO/search-adv.htm

➕ The USPTO website allows users to define a flexible query to pinpoint research interest.

➖ USPTO patent results are difficult to manipulate (large queries result in several pages of patent detail, with a maximum of 50 per page)

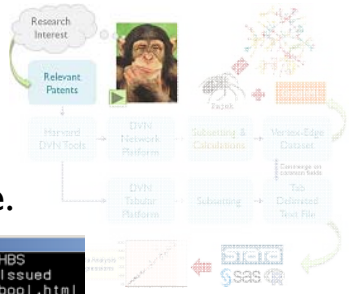
Research Interest



- Our team has developed a solution that allows the flexibility of the USPTO query feature while reducing the redundancies of the USPTO patent result page.

- Let us illustrate our solution through searching for our stated research interest, RNAi.

Research Interest



- Grab patcrawl.zip from the Dataverse.

1

```
Advanced Patent Number Scraper | Developed by HBS
US Patent & Trandmark Office : Patent Search, Issued
 - http://patft.uspto.gov/netahtml/PTO/search-bool.html

 Field codes
 - http://patft.uspto.gov/netahtml/PTO/search-adv.htm

Patent Query (eg. abst/RNAi): abst/RNAi
  > initializing query
  > parsing patent numbers (36 patents in 1 pgs)

Export Options (1=CSV File, 2=On Screen/text file): 1
Filename (default. Pat_Crawl.csv):
  > Pat_Crawl.csv successfully exported

Thanks!  Press Enter to quit the application_
```

Includes:
- Patent Number
- Description and
- Web Link

2

```
Advanced Patent Number Scraper | Developed by HBS
US Patent & Trandmark Office : Patent Search, Issued
 - http://patft.uspto.gov/netahtml/PTO/search-bool.html

 Field codes
 - http://patft.uspto.gov/netahtml/PTO/search-adv.htm

Patent Query (eg. abst/RNAi): abst/RNAi
  > initializing query
  > parsing patent numbers (36 patents in 1 pgs)

Export Options (1=CSV File, 2=On Screen/text file): 2
Please input the following string in the DVN Interface

patent %in% c(6806066, 7022828, 7056704, 7078196, 708743
1991, 7109173, 7129223, 7150970, 7176304, 7232808, 72825
13870, 7419779, 7422853, 7429656, 7459547, 7479369, 747
498316, 7504492, 7514548, 7517864, 7521056, 7521224, 75
7550143, 7550580)
  > Pat_Crawl.txt generated.

Thanks!  Press Enter to quit the application
```

(36 patents in 1 pgs)

Dataverse Network™ PROJECT

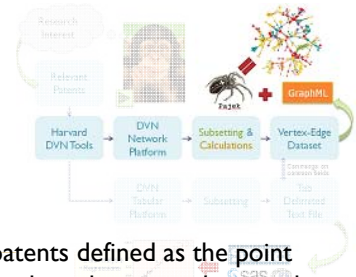
DVN friendly output (text file included).

Note: The .exe version of this script only runs on Windows. To run on other platforms:
1. Install Python.
2. Install PyCurl and BeautifulSoup modules.
3. Execute pat_crawl.py.

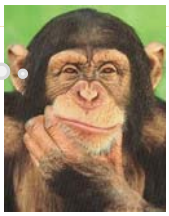
DVN Patent Network Database Project

STEP2 NETWORK PLATFORM

Network Roadmap



RNAi



- With RNAi patents defined as the point of entry, let's explore the network around the time of discovery (1998).
- This is a hands-on example of a workflow that you might follow.

patcrawl.exe

DVN Network Platform

Dataverse Network PROJECT

Relevant Patents → RNAi Inventors → RNAi Network → Localized RNAi Network

Patent Query | Neighborhood | AppYear Query

Subsetting & Calculations

Patent Query for Inventors

RNAI PRESENTATION
DATA FILE: RNAI3_NEW.XML

Back to Study

Paste patent query here.

Write a Manual Query to subset this network data by filtering on vertex or edge attributes.

Manual Query

Attribute Set: Vertex / Edge

Query Workspace:
patent %in% c("7550580", "7550143", "7541344", "7534878", "7527936", "7521224", "7521056", "7517864", "7514548", "7504492", "7498316", "7485414", "7482130", "7479584", "7479369", "7459547", "7429656",

Run

Patent data are stored in edges.

Eliminate disconnected vertices

Drop non-RNAi inventors.

Manual Query Builder

Use these attribute and operator selectors to build a query string in the Query Workspace.

And / Or

Attributes: name, textname, loc, zipcode, city

Operators: Equal to, ==

Values

Add to Query

Subsetting & Calculations

Patent Query Result

- Run the query and check the Subset History.

Subset History

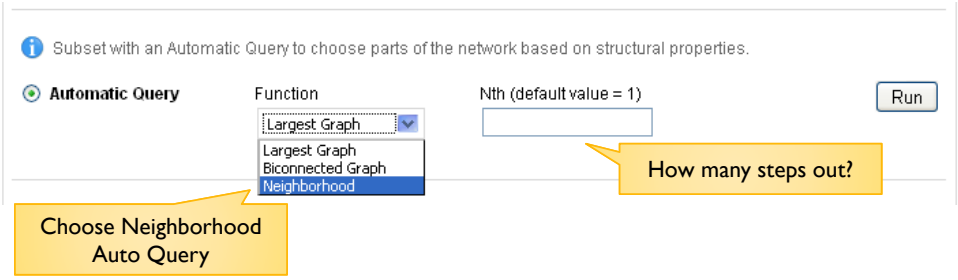
Label	Attribute Set	Query	Vertices	Edges	
Initial State			7308	96501	Restart
Manual Query	edge	patent %in% c("7550580", "7550143", "7541344", "7534878", "7527936", "7521224", "7521056", "7517864", "7514548", "7504492", "7498316", "7485414", "7482130", "7479584", "7479369", "7459547", "7429656", "7422853", "7419779", "7413870", "7368559", "7368248", "7282564", "7232808", "7176304", "7150970", "7129223", "7109173", "7101991", "7094945", "7090990", "7087433", "7078196", "7056704", "7022828", "6806066")	52	96	Undo

Download the last subset.
Download Results

- Whittled down to 52 inventors.
- Now explore their social networks.

Subsetting & Calculations

Neighborhood Query for Networks



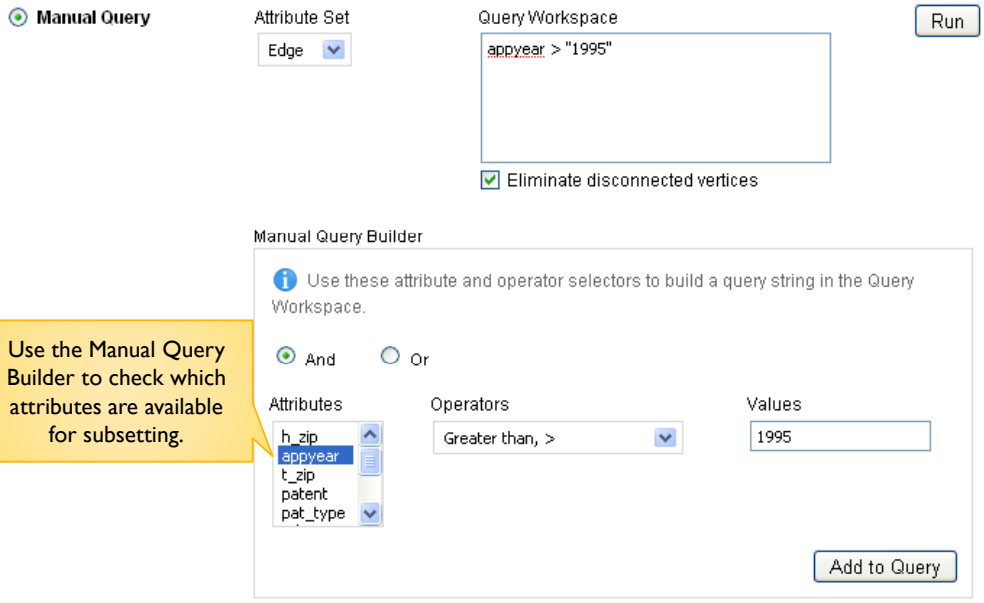
Neighborhood 1		Neighborhood 2	
	Vertices: 222 Edges: 1861		Vertices: 1210 Edges: 12269

Note: N(g, 2) ≡ N(N(g, 1), 1)

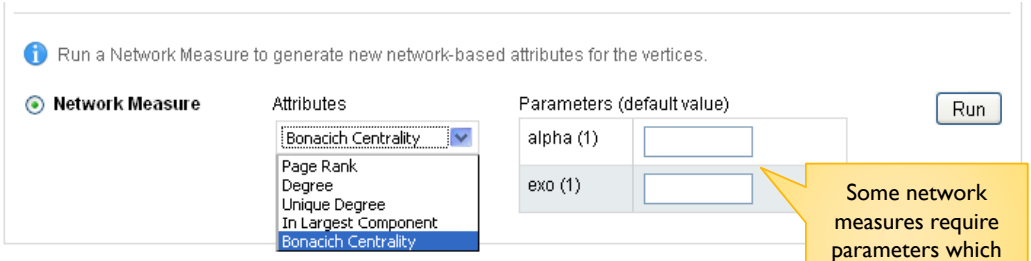
Subsetting & Calculations

AppYear Query to Localize

- Now, drill down to the network surrounding1998, when Fire, et al. published their Nobel Prize winning study on RNAi in *Nature*.



Network Measures



- Measures are dynamically calculated
- Added as a new vertex attribute

Technical Note: Bonacich Centrality measures are only valid for the largest component of the network. Run "Largest Graph" with Nth=1 to extract the largest component before running Bonacich Centrality. For mutiple components, use PageRank. Bonacich centrality will run on multiple components, but its asymptotic behavior is undefined.

Technical Note: Bonacich Centrality is Alpha Centrality, published by Bonacich, et al in 2001. It is parametrized by alpha, the decay parameter, and exo, the exogenous status parameter, which at this point can only be uniform. Alpha can be no larger than the inverse of lambda, the largest eigenvector of the adjacency matrix. The parameter here is a proportion of the inverse of lambda, i.e. alpha = 0.5 corresponds to alpha = 0.5 * 1/lambda.

Export

Soon, this history will be included in download to make analyses easier to replicate.

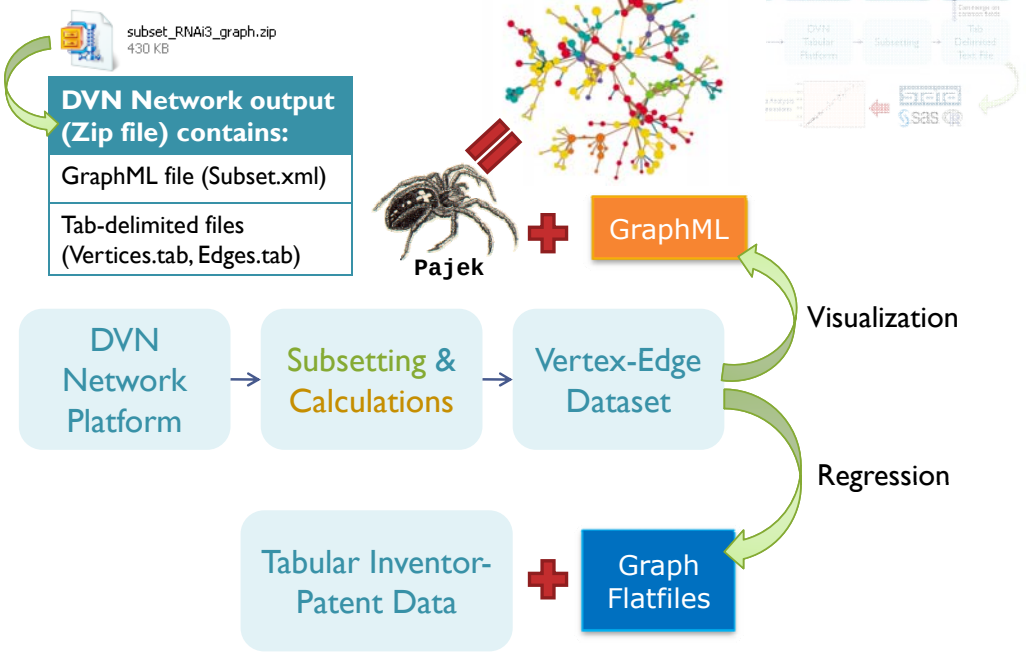
Subset History

Label	Attribute Set	Query	Vertices	Edges	
Initial State			7308	96501	Restart
Manual Query	edge	patent %in% c("7550580", "7550143", "7541344", "7534878", "7527936", "7521224", "7521056", "7517864", "7514548", "7504492", "7498316", "7485414", "7482130", "7479584", "7479369", "7459547", "7429656", "7422853", "7419779", "7413870", "7368559", "7368248", "7282564", "7232808", "7176304", "7150970", "7129223", "7109173", "7101991", "7094945", "7090990", "7087433", "7078196", "7056704", "7022828", "6806066")	52	96	
Automatic Query	N/A	Neighborhood (1)	222	1861	
Automatic Query	N/A	Neighborhood (1)	1210	12269	
Manual Query	edge	appyear > "1995"	1022	9399	
Network Measure	N/A	Unique Degree ()	1022	9399	Undo

Download the last subset.
Download Results

Click to download resulting graph in GraphML and flat file formats (vertex table and edge list).

Output



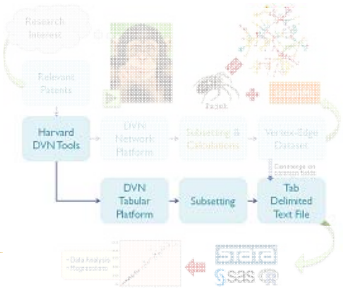
Visualization

- Pajek
  - DVN exports to GraphML, so conversion is necessary.
  - Grab graphml2pajek.zip from our Dataverse.
  - The DVN will soon support direct Pajek downloads.
- Visualization Programs that support GraphML
  - yEd http://www.yworks.com/en/products_yed_about.html
  - GUESS http://graphexploration.cond.org/index.html
  - Gephi http://www.gephi.org
  - Most Java based engines/libraries
- Most graph libraries (igraph, JUNG) parse GraphML.
- Download the R workspace for igraph directly.

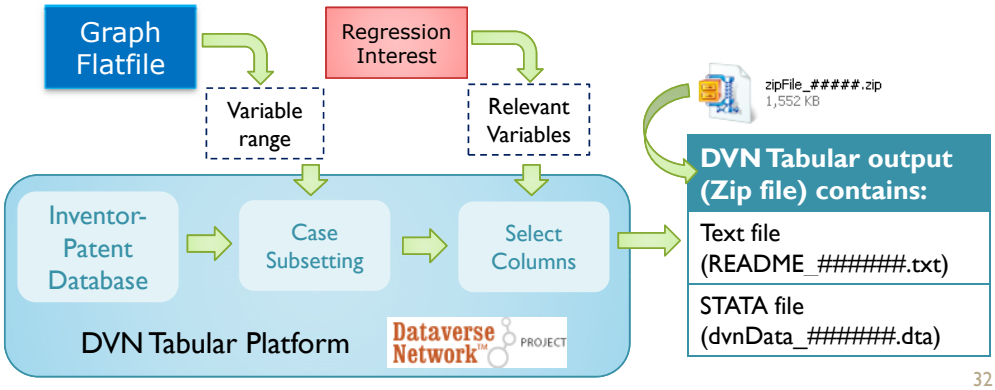
DVN Patent Network Database Project

STEP3 TABULAR PLATFORM

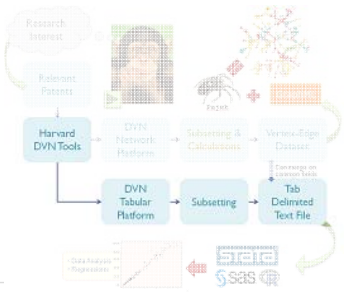
Tabular Platform



- RNAi_inventor_pats.tab
- Subsetting functionality is limited.
- Only download columns you need.
- For additional subsetting, leverage vertex-edge data to constrain variables.
- The DVN team will be working on new features to enhance flexibility for its tabular platform.



Column Download



RNAI PRESENTATION
DATA FILE: RNAI3_INVENTOR_PATS.TAB

Download Subset | Recode & Case-Subset | Descriptive Statistics | Advanced Statistical Analysis

Click the "Download Subset" tab to access this page.

Only check columns you need to reduce memory requirements.

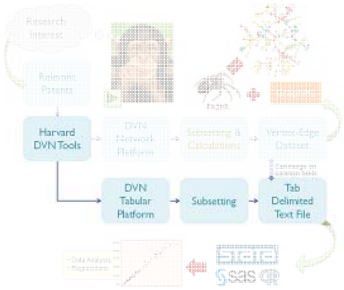
Be sure to show all variables.

Show All
10 Variables
20 Variables
50 Variables
All

Variable Information Table

	Type	Name	Label	Summary
☐	Discrete	invpatseq	invpatseq	
☑	Character	invnum_N	invnum_N	
☐	Character	pat_type	pat_type	
☑	Character	patent	patent	
☑	Character	name	name	
☐	Discrete	inv_seq	inv_seq	
☑	Character	appdate	appdate	

Case Subsetting



RNAI PRESENTATION
DATA FILE: RNAI3_INVENTOR_PATS.TAB

Download Subset | Recode & Case-Subset | Descriptive Statistics | Advanced Statistical Analysis

Click the "Recode & Case Subset" tab to access this page.

See Help Text for subsetting notation.

Selected Variables

invnum_N
patent
name
appdate
USFlg

Start

Use "Drop" and conditions to shrink the dataset.

Choose a variable below, then click here to subset on it.

How to enter a value or range as a condition:

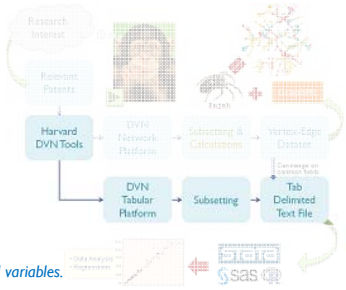
Drop	USFlg		New Value	New Value Label
☑	0	→	0	
☐	1	→	1	
☐	.	→	.	

Add Value/Range

Apply Recodes

Apply recodes when done, then return to Download page.

Tabular Platform



Inventor-Patent Database

*HBS algorithm generated variables.

INVENTOR	
Invnum_N	*Unique inventor number.*
Name	Inventor name (Last, first middle initial)
Inv_Seq	Inventor number on patent.

PATENT	
Invpatseq	*Generated patent history sequence for inventor.*
Pat_Type	94% Utility (U), 5% Design (D).
Patent	USPTO assigned patent number.
AppDate	Patent application date.
GDate	Patent grant date.
Claims	Total patent claims.

INVENTOR LOCATION	
City	Inventor's city.
Loc	(US only) State or Country code.
Zipcode	(US only) Zipcode.
Usflg	US = 1 (50 states, D.C. and territories)

ASSIGNEE	
Assignee	Primary firm associated with patent.
Numasg	*Generated assignee number.*

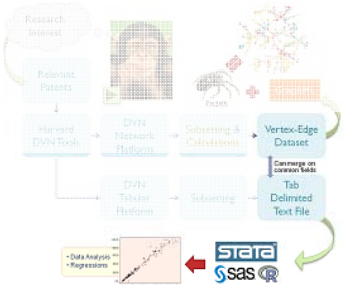
CLASSES	
Classes	Classes separated by "|". Main-Sub. Primary class listed first. 7 maximum.
Scls_cnt	*Count of total classes.*
Scls_1	*Flag. Contains a first occurrence of class.*
Scls_pair_1	*Total of first occurence of class-pairs.*

CITATION (Fieldnames combine the concepts)	
Back	*Backward citations (patents cited to)*
Forw	*Forward citation (patents cited from).*
5	*5 year window.(otherwise, for all time).*
SC	*Assignee self citation.*

DVN Patent Network Database Project

STEP4 REGRESSIONS

Regressions



- Researchers can easily run regressions using the DVN network or tabular data (flat files).

☆ Suppose you wanted to merge the two datasets.

Vertex-Edge Dataset

Can merge on common fields

Tab Delimited Text File

SQL Code

```
select *
from vertices as a
        inner join
     dvnData as b
on a.invnum_N = b.invnum_N
```

STATA Code

See next slide…

Regressions, Stata import

Grab vertex_dta_merge.do from our Dataverse.

```
clear
*adjust to your personal directory
cd "C:\[[path to data]]"

*convert vertices to dta format
    insheet using "vertices.tab"
    generate invnum_N=string(invnum_n, "%15.0g")
    drop invnum_n
    save "vertices", replace

*sort dvn file
    use [[dvnDataFile]], clear
    sort invnum_N
    save [[dvnDataFile]], replace

*merge [[dvnDataFile]] with records in vertices
    use "vertices.dta", clear
    sort invnum_N
    merge invnum_N using [[dvnDataFile]]

*tidy up and save
    drop if _merge != 3
    drop _merge
    save "[[merged file]]"
```

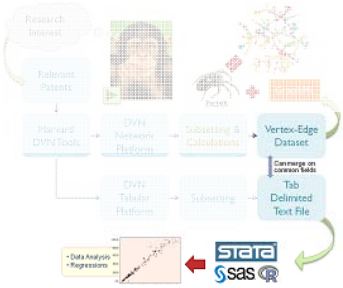
Change working directory to where vertices.tab and dvnDataFile_#######.dta were extracted.

STATA interprets invnum_N as a numerical value, even though it is stored as string. We need to convert it back to string, retaining all digits, to merge.

Change this to the name of the dta file from the DVN Tabular output.

Change this to what you want to name the final merged dataset.

Regressions, Examples



Example:

```
nbreg forw bonacich_c claims back5
```

```
Fitting Poisson model:

Iteration 0:   log likelihood = -11763.919
Iteration 1:   log likelihood =  -11763.71
Iteration 2:   log likelihood =  -11763.71

Fitting constant-only model:

Iteration 0:   log likelihood = -8095.5749  (not concave)
Iteration 1:   log likelihood = -5557.1525
Iteration 2:   log likelihood = -5437.2324
Iteration 3:   log likelihood = -5433.9893
Iteration 4:   log likelihood = -5433.9887
Iteration 5:   log likelihood = -5433.9887

Fitting full model:

Iteration 0:   log likelihood = -5426.2683
Iteration 1:   log likelihood = -5424.9179
Iteration 2:   log likelihood = -5424.9137
Iteration 3:   log likelihood = -5424.9137

Negative binomial regression                    Number of obs   =      12357
                                                LR chi2(3)      =      18.15
Dispersion     = mean                           Prob > chi2     =     0.0004
Log likelihood = -5424.9137                     Pseudo R2       =     0.0017

------------------------------------------------------------------------------
        forw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
bonacich_c~1 |   1.257425   1.456492     0.86   0.388    -1.597246    4.112097
      claims |   .0110707   .0029924     3.70   0.000     .0052056    .0169357
       back5 |   .0081291   .0063491     1.28   0.200    -.0043149    .0205732
       _cons |  -1.573448   .0766626   -20.52   0.000    -1.723704   -1.423192
-------------+----------------------------------------------------------------
     /lnalpha |   3.117988   .0423964                      3.034893    3.201084
-------------+----------------------------------------------------------------
        alpha |   22.60087   .9581962                      20.79875    24.55913
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) = 1.3e+04 Prob>=chibar2 = 0.000
```


Questions?