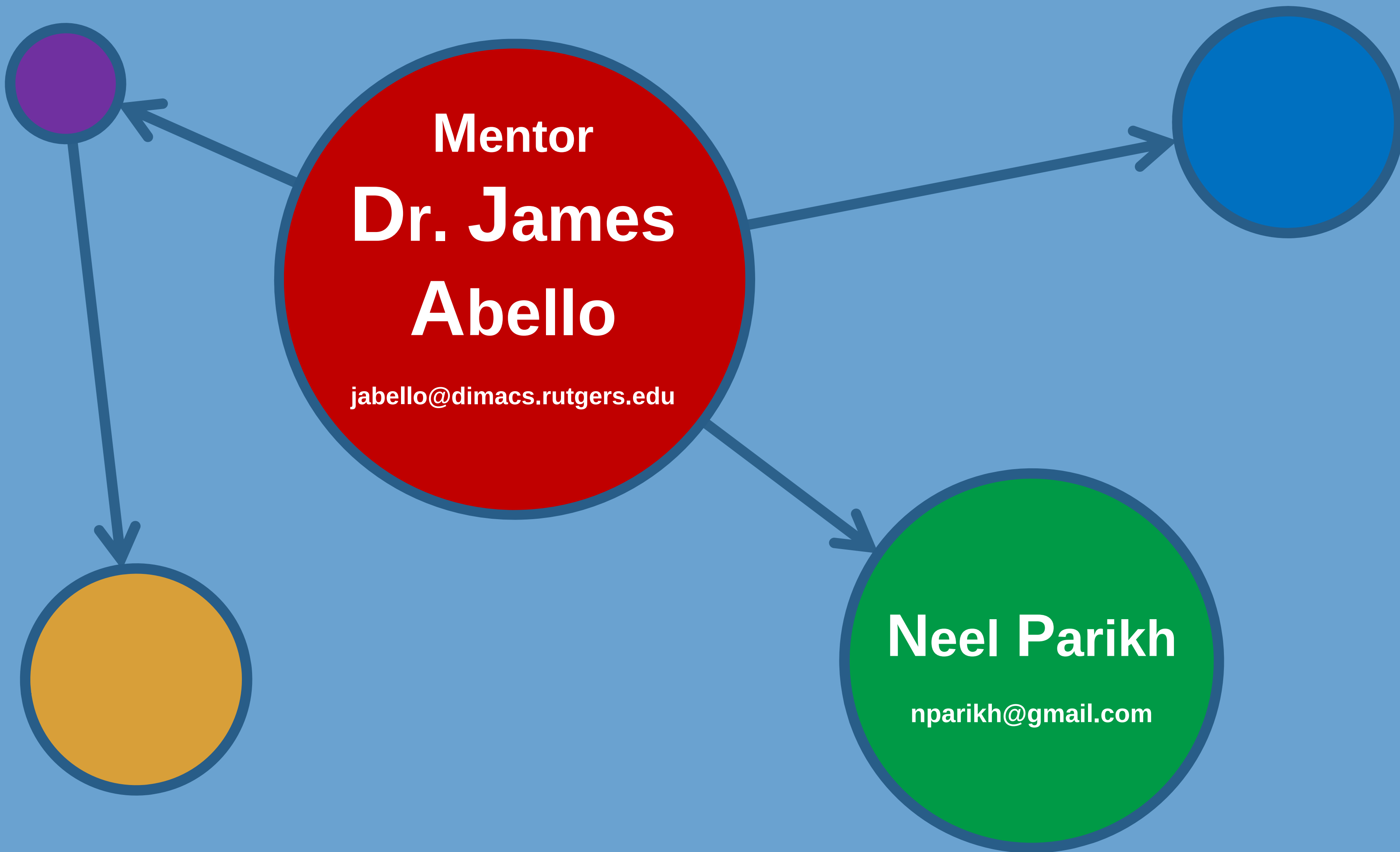
REU 2012

Visualizing of Time-Varying Graphs Part II

July 19, 2012

PROJECT TEAM



Problem Statement

Visualizing Dynamic Graphs

- Data is being collected in a streaming fashion
- The graph structure is changing over time
- Data attributes are being updated
    - e.g. edge weight, statistical measures, etc.

Challenges

- Positioning vertices to display graph clearly on screen
- Preserving viewers mental map of graph
- Conveying underlying semantics

GraphStream – A Dynamic Graph Library

Designed to Represent Dynamic Graphs

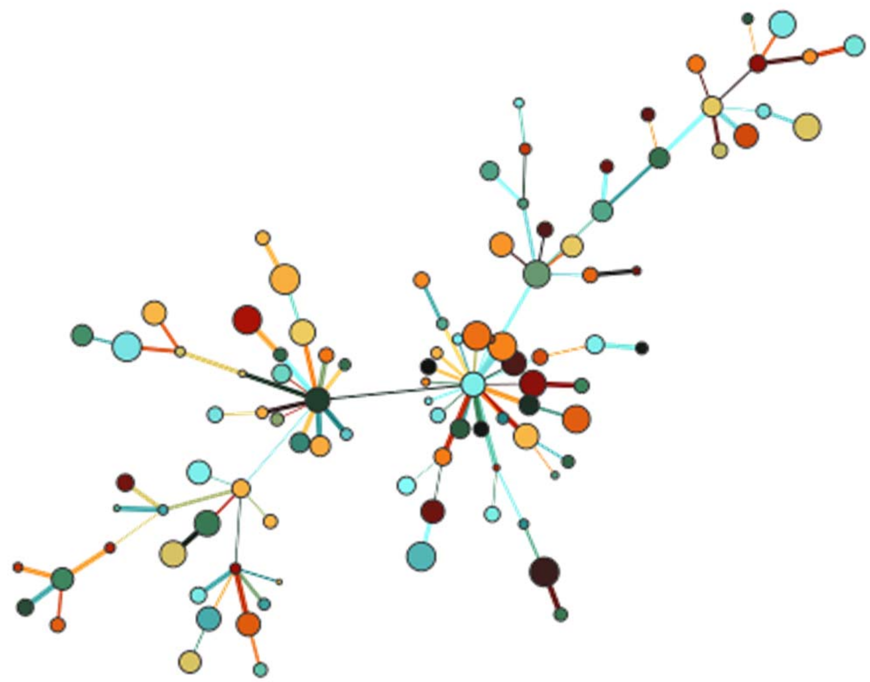
- Updateable graph attributes
- Display applies force-directed layout
- Allows user interaction

Uses Event Streaming Model

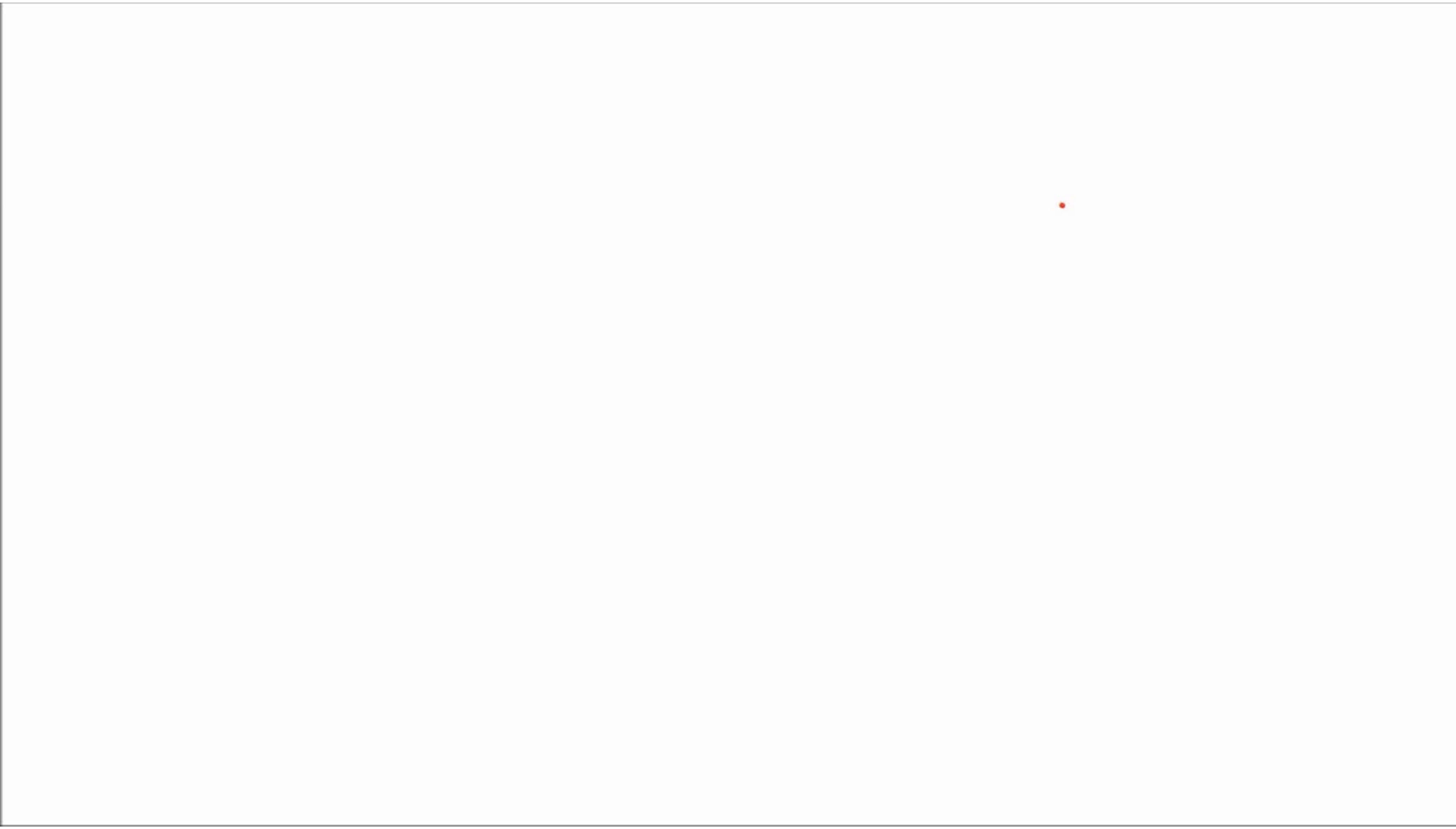
- Discrete add, delete or change events for nodes, edges and the overall graph
- Events can be streamed to display or file

Visual Attributes Defined by CSS

- Dynamically assigns visual styles
- Dramatically reduces number events needed to update visualization



Streaming A Binary Tree With FD Layout



Dynamically Visualizing Graph Algorithms

Generation of a Random Weighted Graph

- Addition of nodes and randomly weighted edges
- Uses the Dorogovstev-Mendes algorithm

Calculate the Minimum Spanning Tree

- By Kruskal MST algorithm
- Colors non-tree edges red

Remove Non-Tree Edges

- Sends remove events for each non-tree edge
- Force directed layout positions the resulting tree

Dynamically Visualizing Graph Algorithms

Data Processing Pipeline

A Set of Programs That Process Data Sequentially

- Given raw input, generate a time varying graph representation
- Each program is a single step in that process

The Goal Was To Streamline and Unify the Process

- The output of one process is the input to the next
- Fully automate the pipeline
- Remain general enough to be used with any streaming data

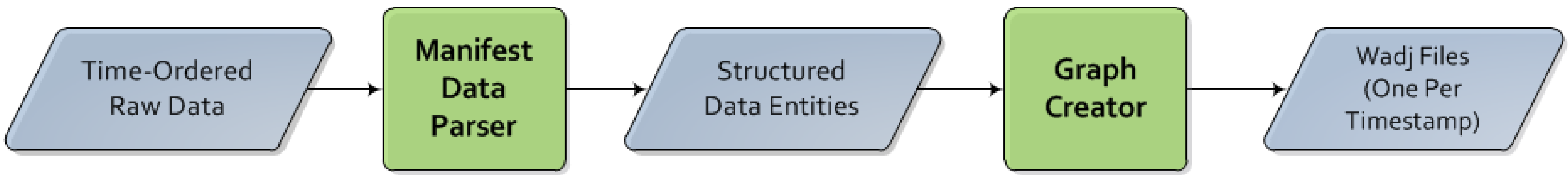
Three Main Components

- 1. Data Extraction
- 2. Statistical Analysis
- 3. Visualization Rendering

Data Extraction

Raw Data in the ACSFOI Format

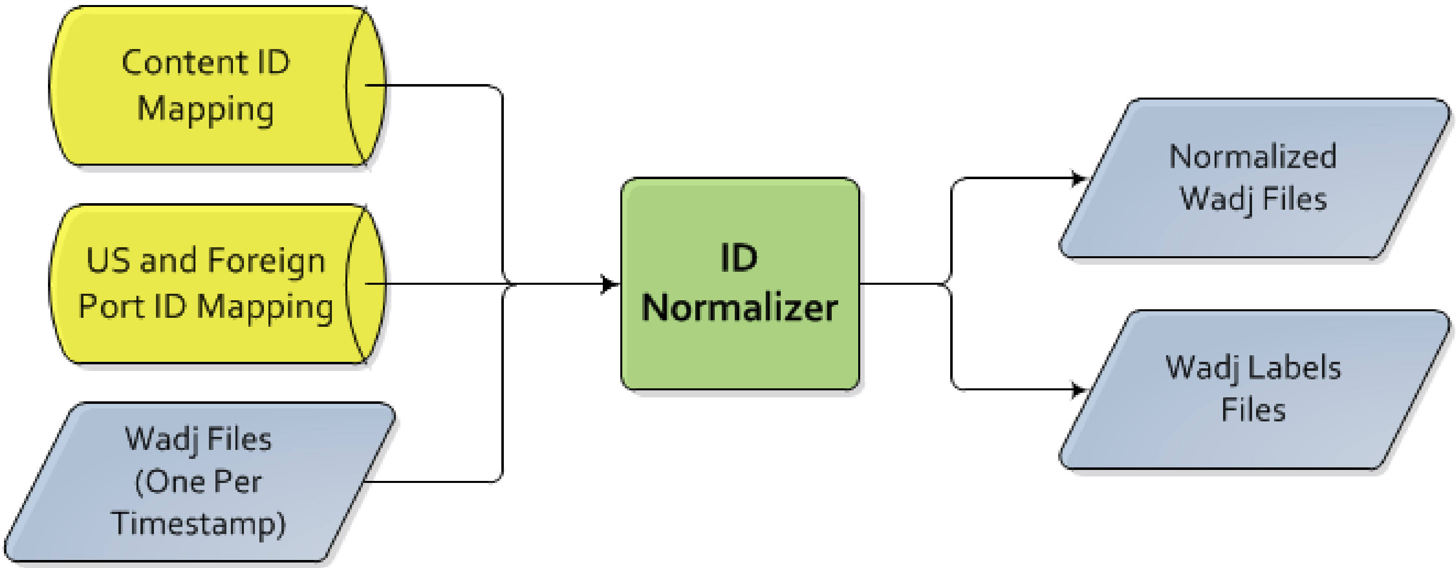
```
1LNCFNOMSC PERU               735  2709090828LNCF2101246571  4753100000000973CTN
00000027000KG00000000000CMDROCCA DI SAN CASC         8619053001100337550
20090828HMSCUS8254472
1MSCU1344460   2887788                      G002000000008060000080022G0LHH
2PORCELLANA DI ROCCA SPA           VIA NAZIONALE 43                  ROCCA DI SAN CASCIAN, FO  47017  IT
0546-652111
3EMSER TILE LLC                    8431 SANTA MONICA BLVD.           LOS ANGELES, CA  90069  US
0013236502000
5MSCU1344460   0000000973CERAMIC TILES                                         000000000000000000
7MSCU1344460   EMSER TILE LLC
```



Converts Raw Data to Graphs in Wadj Format

- Specific to the type of raw data
- The only data step

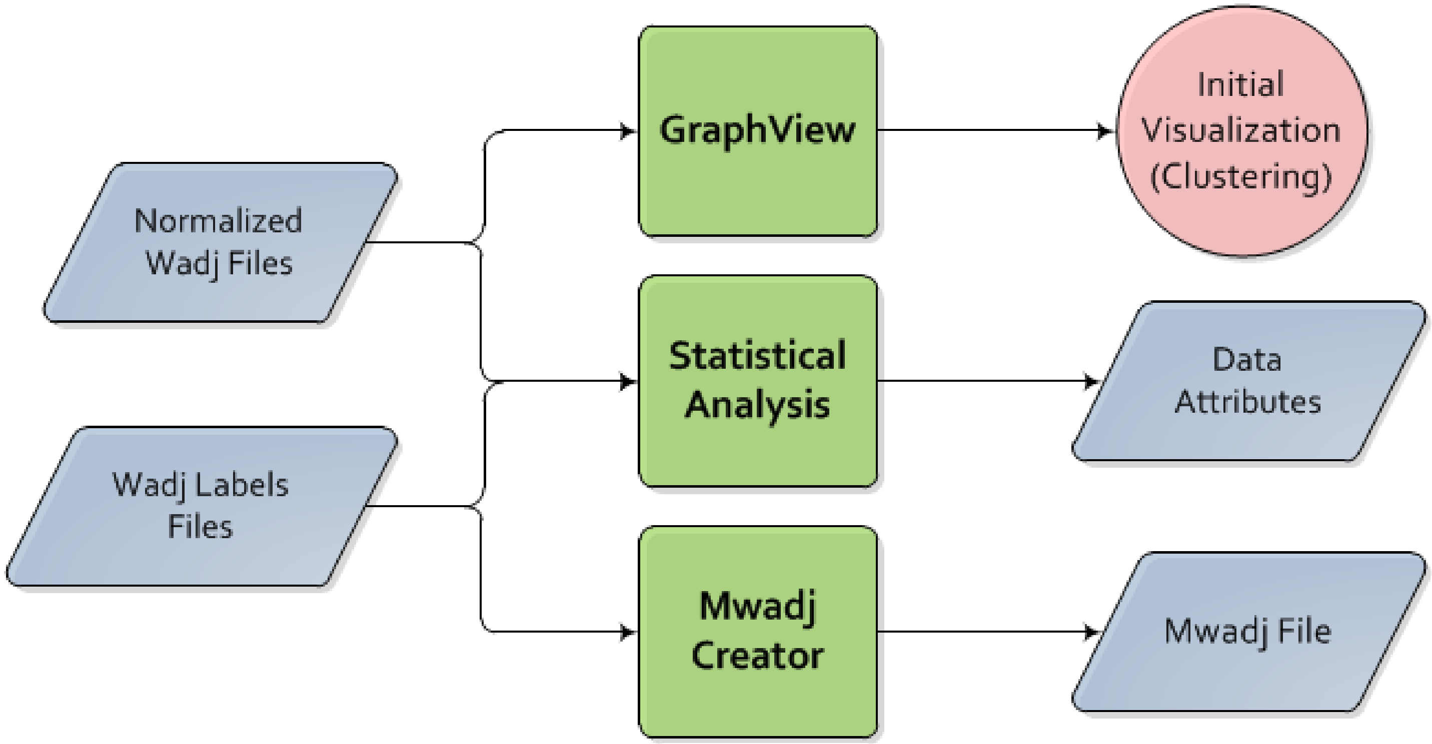
Data Extraction



Fixes Critical Bug In DataParser

- Makes content ID codes consistent
- Generates labels for each vertex

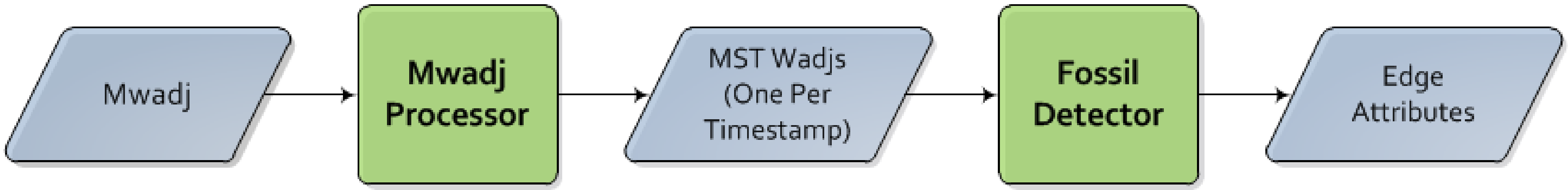
Statistical Analysis



Generates Data Attributes

- Calculation of firing rate, recency, normalized frequency
- Will be the values used in dynamic visualization

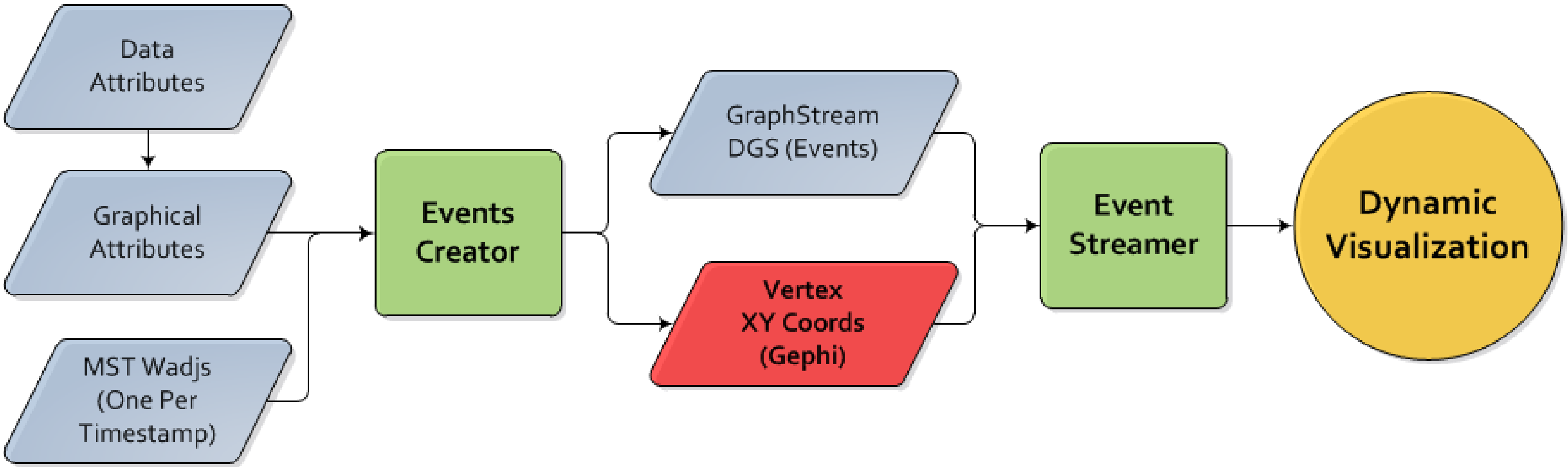
Statistical Analysis



Edges Weighted By Discrepancy

- Statistical measure that ranks each edge by how much it deviates from the behavior of the system
- Creates cumulative maximum spanning trees for each timestamp
- Fossil edges are the most discrepant for a particular timestamp

Visualization Rendering



Events Generated and Streamed to Display

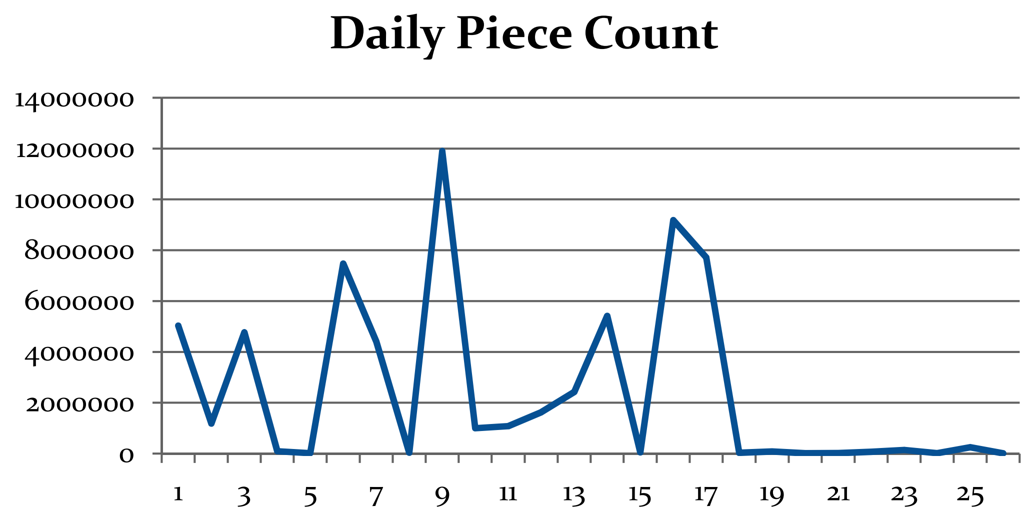
- Analysis of changes between the graph during each timestamp
- Graphical attributes calculated
- Coordinates still determined manually

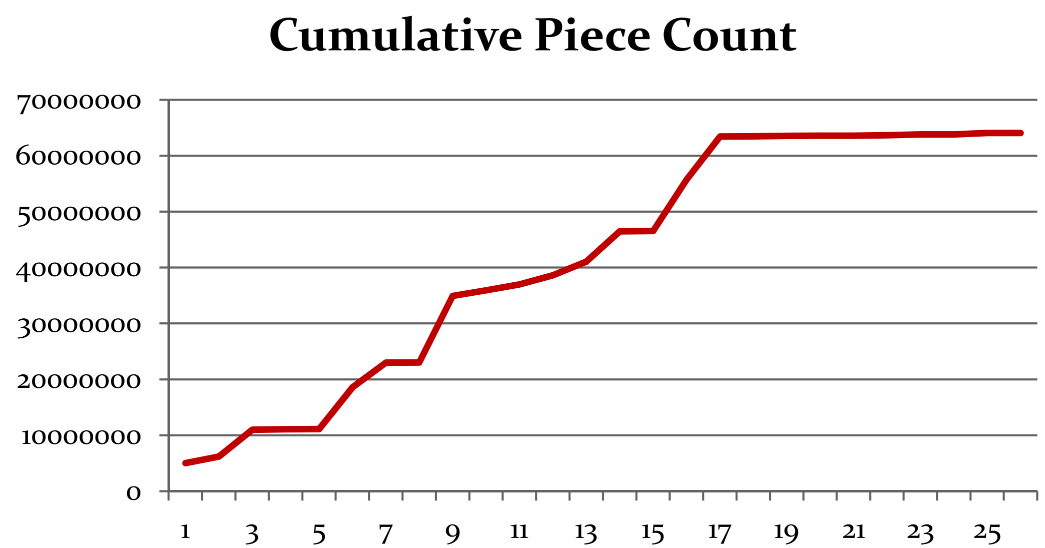
Japan Tsunami Manifest Data

- Data spans time period before and after Japan tsunami on March 11, 2011
  - 26 Days: September 15, 2010 to August 31, 2011
  - Wednesdays separated by two weeks
- Graph Representation
  - Vertices
    - Port Pairs – ex. TokyoJapan-> LosAngeles-CA
    - Contents – ex. HomeGarden-KitchenDining-Tableware
  - Edges
    - Weighted by discrepancy in the quantity of particular goods shipped
    - Cumulative measure over entire time period of data
    - Anomalous “Fossil” edges are the most discrepant
- Visual representation of those ports within 400 miles of Sendai, Japan that exhibited the most anomalous traffic pattern after the Tsunami.
  - Sendai was struck most directly by tsunami
  - Shows how traffic patterns of nearby ports were affected

Shipment Volume From Kobe, Japan

Day Num	Volume	Cumulative Vol.
1	5038460	5038460
2	1180807	6219267
3	4774311	10993578
4	93084	11086662
5	20707	11107369
6	7477221	18584590
7	4407894	22992484
8	36417	23028901
9	11900837	34929738
10	999279	35929017
11	1083538	37012555
12	1628028	38640583
13	2422466	41063049
14	5417628	46480677
15	53122	46533799
16	9184428	55718227
17	7717955	63436182
18	33214	63469396
19	84282	63553678
20	13448	63567126
21	17990	63585116
22	70128	63655244
23	143202	63798446
24	16916	63815362
25	249134	64064496
26	11828	64076324

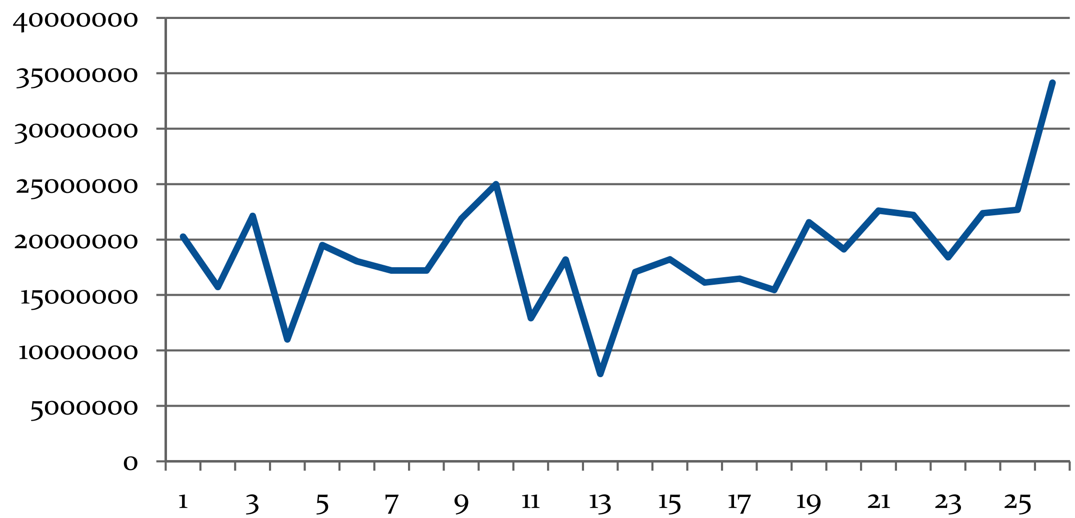




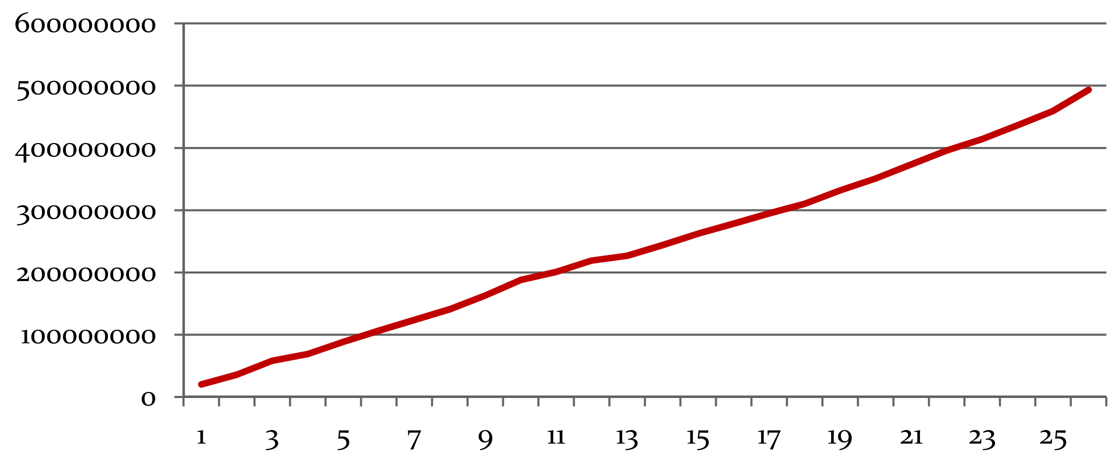
Shipment Volume from Chinese Ports

Day Num	Volume	Cumulative Vol.
1	20274897	20274897
2	15724426	35999323
3	22140936	58140259
4	11000860	69141119
5	19508625	88649744
6	18051123	106700867
7	17206201	123907068
8	17210143	141117211
9	21884277	163001488
10	24991930	187993418
11	12912541	200905959
12	18208131	219114090
13	7884810	226998900
14	17083055	244081955
15	18223877	262305832
16	16125343	278431175
17	16483126	294914301
18	15459485	310373786
19	21563383	331937169
20	19128281	351065450
21	22615280	373680730
22	22224030	395904760
23	18415951	414320711
24	22386592	436707303
25	22685855	459393158
26	34150591	493543749

Daily Piece Count

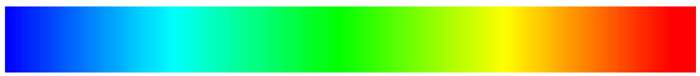


Cumulative Piece Count



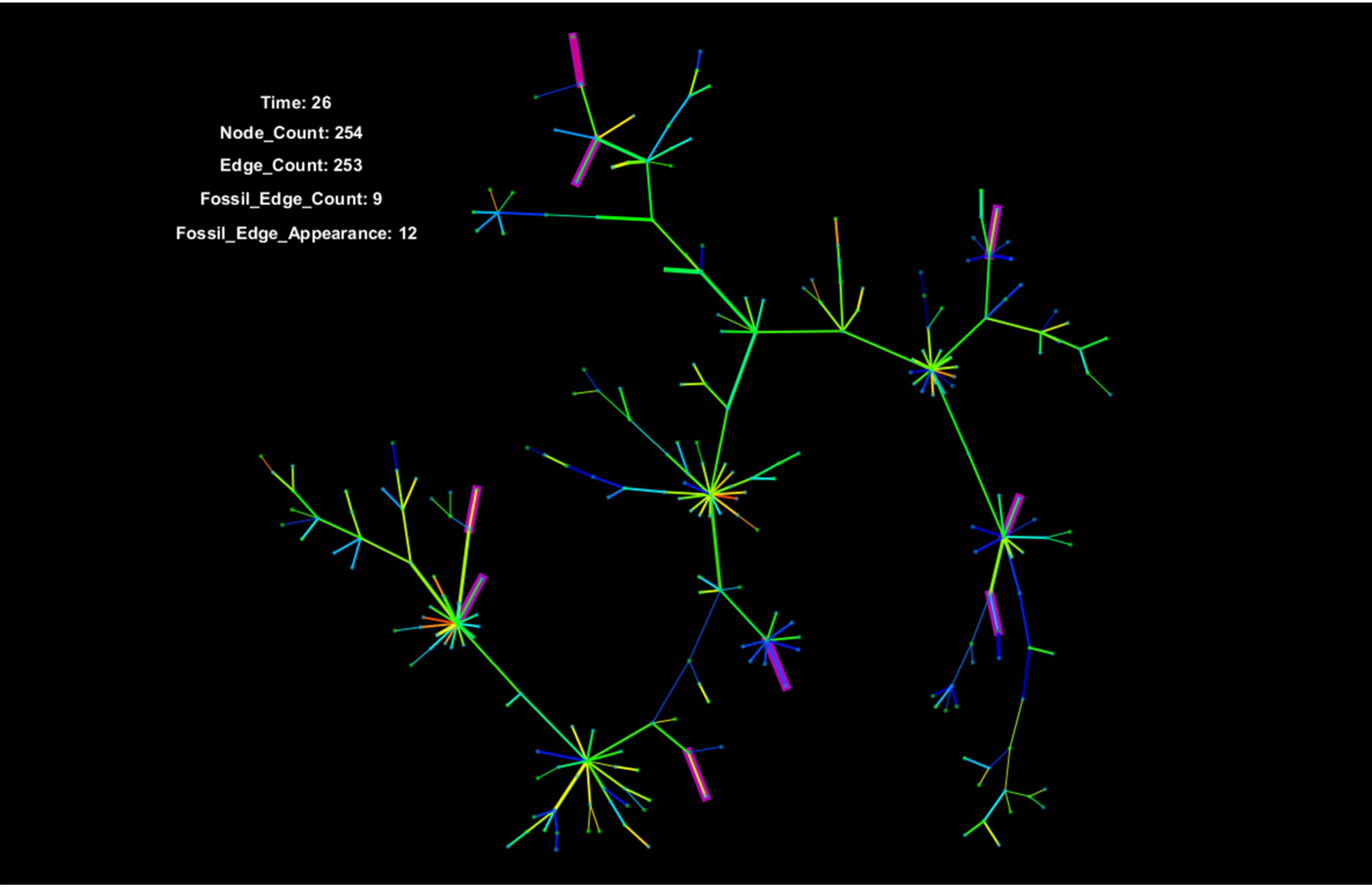
Assigning Visual Attributes

- Two types of vertices
  - **Green** nodes are port-pairs
  - **Blue** nodes are specific content categories

- Visualization Measures
  - Edge discrepancy weight – Cumulative measure of anomaly
    - Colored by linear mapping to the RGB color space
    - **Low**  **High**
  - Fossil edges – Were most anomalous on a particular day
    - Have weight with greatest number of standard deviations from the mean
    - **Highlighted In Magenta**
  - Edge firing rate - $(\sum_{1}^{t} units\ shipped)/t$
    - Edge **Thickness**
    - Firing rate follows a power-law distribution
    - Scaled using the function: $width = \ln\,(FiringRate + e)$

Anomaly in Shipment Patterns Over the Time Period

Ports Near Sendai, Japan



Future Work

Edge Oriented Processing

- Right now inputs and output are data files
- Atomic unit of data should be a triple
  - < Foreign Port, US Port, Content >

Buffering I/O Between Processing Steps

- Eliminate time-consuming i/o with hard disk
- Parallel processing
- Real time visualization

Fully Automated Layout

- Deterministically minimize edge crossings
- Combine Yifan Hu Proportional, Force Atlas and Fruchterman-Reingold

Visual Effects

- Fading in/out of graph elements, panning and zooming, sprites
- Generated in independent threads

Questions and Comments

References

1. Abello J, Eliassi-Rad T, Devanur N (2010), *Detecting Novel Discrepancies in Communications Networks*, International Conference on Data Mining, ICDM 2010: 8-17, Sidney Australia, Dec 2010.

2. Chazelle B (2000), The Discrepancy Method: Randomness and Complexity, Cambridge University Press, New York.

3. Bastion M, Heymann S, Jacomy M (2009), *Gephi: An Open Source Software for Exploring and Manipulating Networks*, Proceedings of the Third International Conference on Weblogs and Social Media, May 2009.

4. Abello J, Chen M, Parikh N (2012), *Time Discrepant Shipments in Manifest data*, Handbook of Operations Research for Homeland Security, Springer, New York.