Topic II.2: Connecting the Dots

Discrete Topics in Data Mining
Universität des Saarlandes, Saarbrücken
Winter Semester 2012/13

T II.2: Connecting the Dots

Shahaf & Guestrin 2010, 2012; Shahaf, Guestrin & Horvitz 2012a

Connecting the Dots

- *What connects two events?*
  - E.g. 2007 housing bubble burst and Obamacare
- More concretely, given two user-selected news articles, find a series of news articles that explain how these articles are connected
  - Each successive article should reasonably connect to the previous one
  - Together, the articles should tell a coherent story
- **Goals**: Formalise "connected" and "coherent" and find the good chains

Shahaf & Guestrin 2010, 2012

Example Chain

B1: Talks Over **Ex-Intern's Testimony On Clinton** Appear to Bog Down

B2: **Clinton Admits Lewinsky** Liaison to Jury; Tells Nation 'It was Wrong,' but Private

B3: G.O.P. Vote Counter in House **Predicts Impeachment of Clinton**

B4: **Clinton Impeached**; He Faces a Senate Trial, 2d in History; Vows to Do Job till Term's 'Last Hour'

B5: **Clinton's Acquittal**; Excerpts: Senators Talk About Their Votes in the Impeachment Trial

B6: Aides Say Clinton Is Angered As **Gore Tries to Break Away**

B7: As **Election Draws Near**, the Race Turns Mean

B8: **Contesting the Vote**: The Overview; Gore asks Public For Patience; Bush Starts Transition Moves

Shahaf & Guestrin 2010

First Idea

- Take the news articles as vertices in the graph
- Add an edge between two vertices if the articles share words
  - Perhaps just titles and/or require multiple instances
    - In general, measure similarity
  - Direction of the edge based on chronological order
- Find the shortest path between the two vertices
  - Breath-first search

An Example of the Simple Idea

A1: Talks **Over Ex-Intern's Testimony** On Clinton Appear to Bog Down

A2: Judge Sides with the Government in **Microsoft Antitrust Trial**

A3: Who will be **the Next Microsoft**?
trading at a **market** capitalization…

A4: Palestinians Planning to Offer **Bonds on Euro. Markets**

A5: Clinton Watches as **Palestinians Vote to Rescind** 1964 Provision

A6: **Contesting the Vote:** The Overview; Gore asks Public For Patience; Bush Starts Transition Moves
The **Clinton** administration has denied…

Shahaf & Guestrin 2010

An Example of the Simple Idea

Court trials

A1: Talks **Over Ex-Intern's Testimony** On Clinton Appear to Bog Down

A2: Judge Sides with the Government in **Microsoft Antitrust Trial**

A3: Who will be **the Next Microsoft**?
trading at a **market** capitalization…

A4: Palestinians Planning to Offer **Bonds on Euro. Markets**

A5: Clinton Watches as **Palestinians Vote to Rescind** 1964 Provision

A6: **Contesting the Vote:** The Overview; Gore asks Public For Patience; Bush Starts Transition Moves
The **Clinton** administration has denied…

Shahaf & Guestrin 2010

An Example of the Simple Idea

Court trials

Microsoft

A1: Talks **Over Ex-Intern's Testimony** On Clinton Appear to Bog Down

A2: Judge Sides with the Government in **Microsoft Antitrust Trial**

A3: Who will be **the Next Microsoft**?
trading at a **market** capitalization…

A4: Palestinians Planning to Offer **Bonds on Euro. Markets**

A5: Clinton Watches as **Palestinians Vote to Rescind** 1964 Provision

A6: **Contesting the Vote:** The Overview; Gore asks Public For Patience; Bush Starts Transition Moves
The **Clinton** administration has denied…

Shahaf & Guestrin 2010

An Example of the Simple Idea

Court trials

Microsoft

Markets

A1: Talks **Over Ex-Intern's Testimony** On Clinton Appear to Bog Down

A2: Judge Sides with the Government in **Microsoft Antitrust Trial**

A3: Who will be **the Next Microsoft**?
trading at a **market** capitalization…

A4: Palestinians Planning to Offer **Bonds on Euro. Markets**

A5: Clinton Watches as **Palestinians Vote to Rescind** 1964 Provision

A6: **Contesting the Vote:** The Overview; Gore asks Public For Patience; Bush Starts Transition Moves
The **Clinton** administration has denied…

Shahaf & Guestrin 2010

An Example of the Simple Idea

Court trials

Microsoft

Markets

Palestinians

A1: Talks **Over Ex-Intern's Testimony** On Clinton Appear to Bog Down

A2: Judge Sides with the Government in **Microsoft Antitrust Trial**

A3: Who will be **the Next Microsoft**?
trading at a **market** capitalization…

A4: Palestinians Planning to Offer **Bonds on Euro. Markets**

A5: Clinton Watches as **Palestinians Vote to Rescind** 1964 Provision

A6: **Contesting the Vote:** The Overview; Gore asks Public For Patience; Bush Starts Transition Moves
The **Clinton** administration has denied…

Shahaf & Guestrin 2010

An Example of the Simple Idea

Court trials

Microsoft

Markets

Palestinians

Votes & Clinton

A1: Talks **Over Ex-Intern's Testimony** On Clinton Appear to Bog Down

A2: Judge Sides with the Government in **Microsoft Antitrust Trial**

A3: Who will be **the Next Microsoft**?
trading at a **market** capitalization…

A4: Palestinians Planning to Offer **Bonds on Euro. Markets**

A5: Clinton Watches as **Palestinians Vote to Rescind** 1964 Provision

A6: **Contesting the Vote:** The Overview; Gore asks Public For Patience; Bush Starts Transition Moves
The **Clinton** administration has denied…

Shahaf & Guestrin 2010

An Example of the Simple Idea

Court trials

Microsoft

Markets

Palestinians

Votes & Clinton

A1: Talks **Over Ex-Intern's Testimony** On Clinton Appear to Bog Down

A2: Judge Sides with the Government in **Microsoft Antitrust Trial**

A3: Who will be **the Next Microsoft**?
trading at a **market** capitalization…

A4: Palestinians Planning to Offer **Bonds on Euro. Markets**

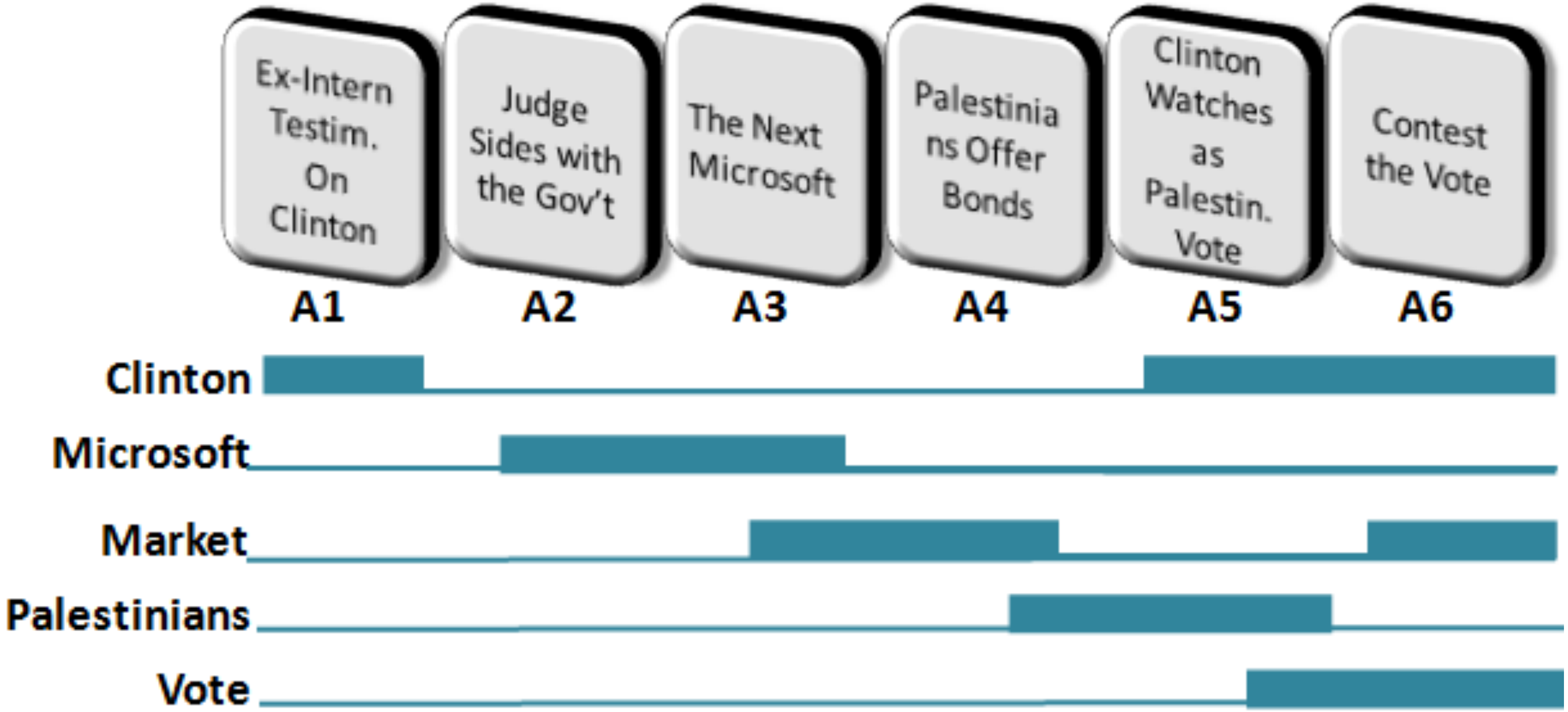
A5: Clinton Watches as **Palestinians Vote to Rescind** 1964 Provision

A6: **Contesting the Vote:** The Overview; Gore asks Public For Patience; Bush Starts Transition Moves
The **Clinton** administration has denied…

Not very coherent

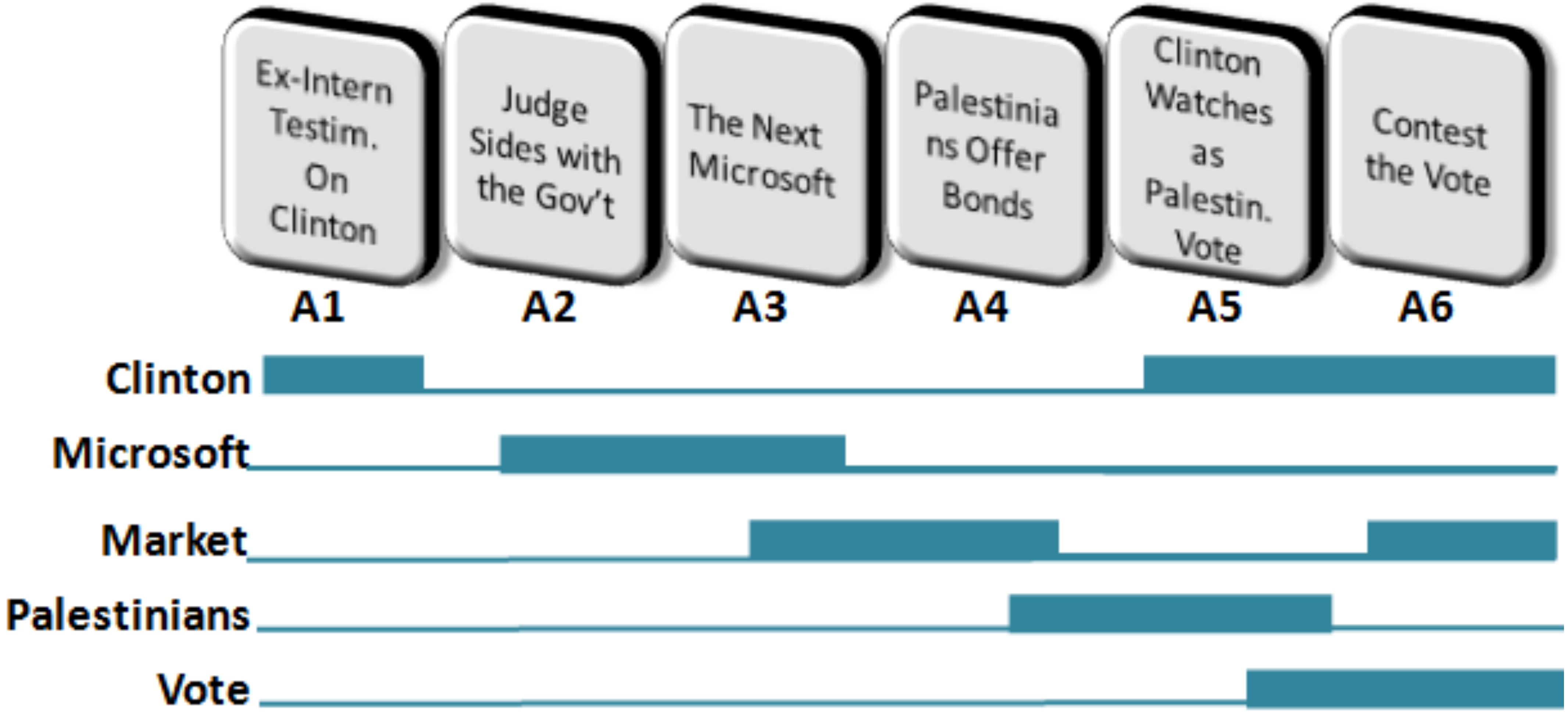
Shahaf & Guestrin 2010

Not-So Coherent Story



Shahaf & Guestrin 2010

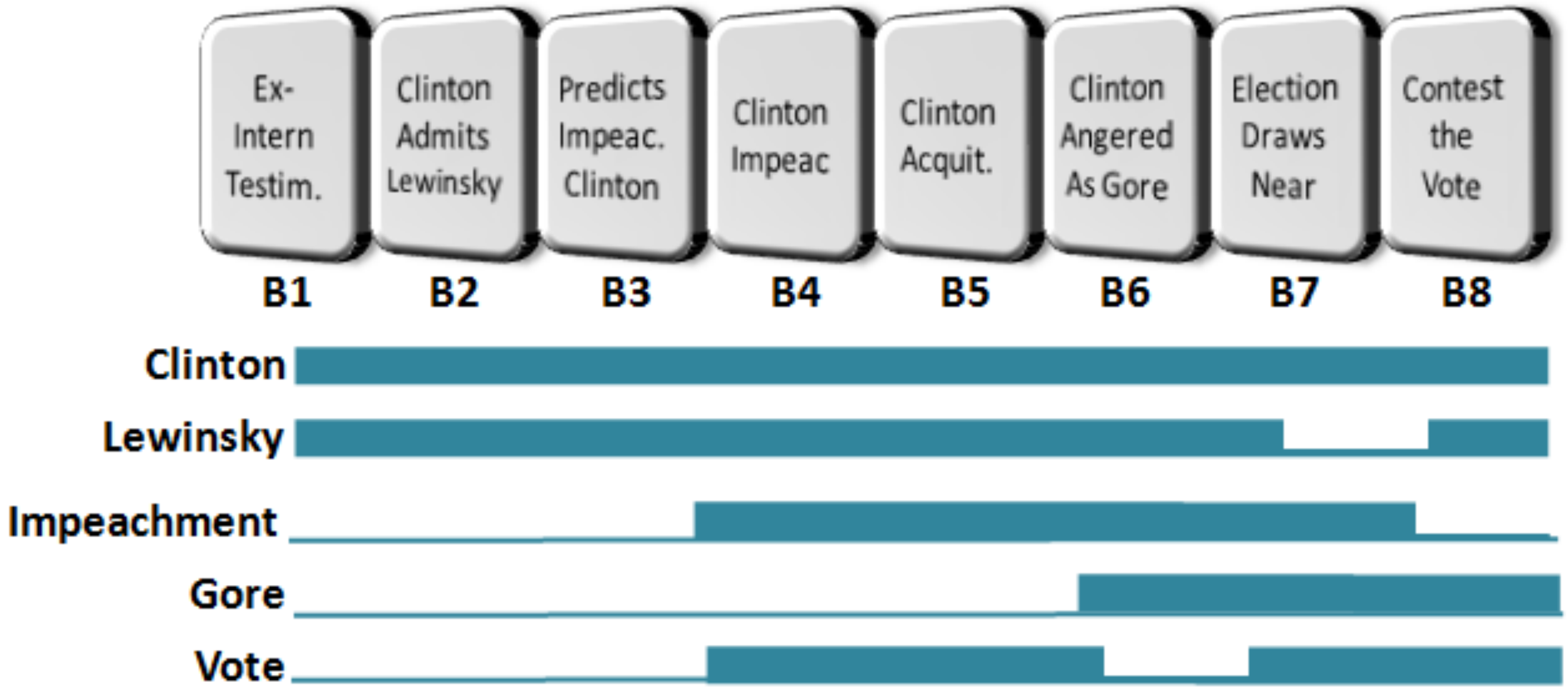
Not-So Coherent Story



Topic changes in every transition

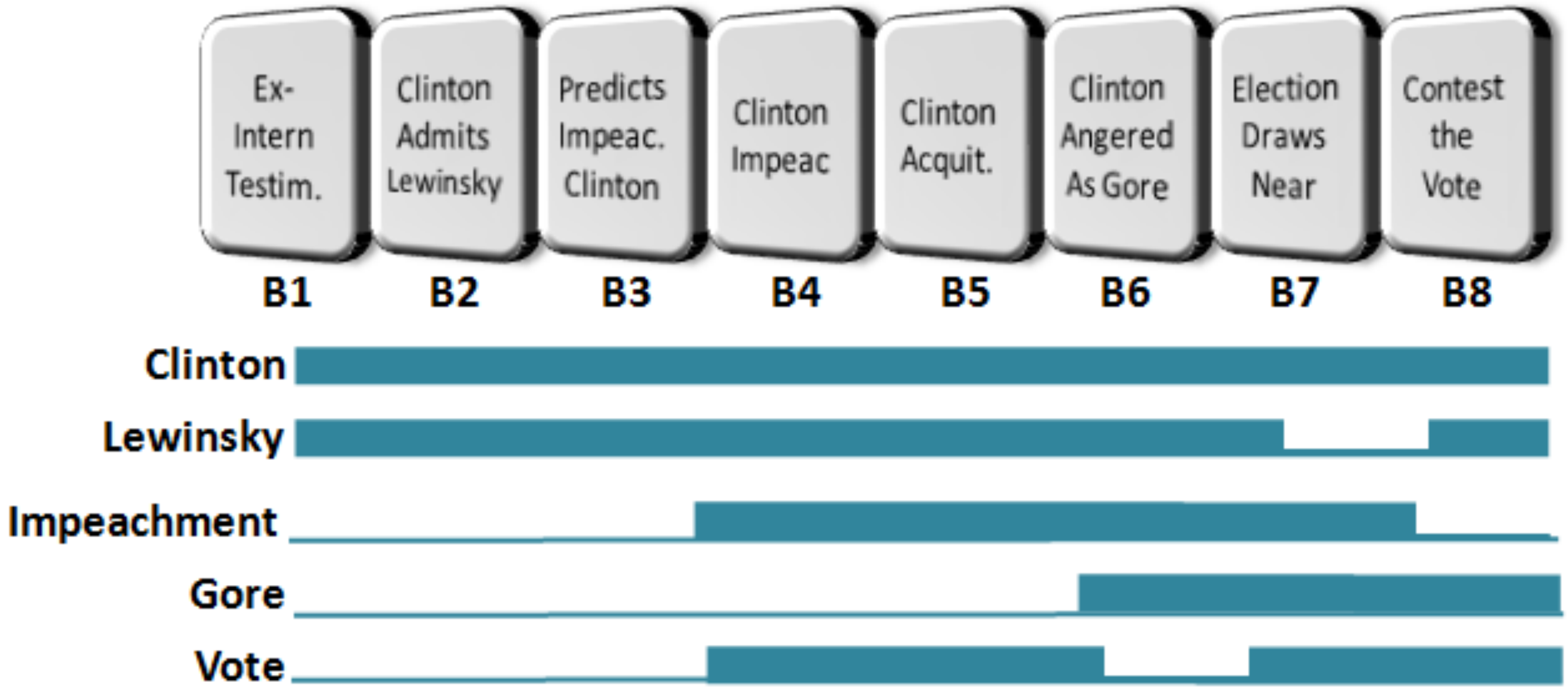
Shahaf & Guestrin 2010

More Coherent Story



Shahaf & Guestrin 2010

More Coherent Story



Topic consistent over transitions

Shahaf & Guestrin 2010

Intuition for a Good Chain

- Every transition must be strong
  - Articles must be well linked
- There must be a global theme
  - Topic that spans (almost) all articles
- No jitteriness
  - No switching topics back-and-forth
- Short

First Attempt on Strong Transitions

- *A chain is as weak as its weakest link*
  - We score the chain by its minimum-strength transition
- First idea for the strength of transition: shared words
- Let d be a document (bag-of-words) and write $w \in d$ if word w appears in document d
  - Let the chain C be $\langle d_1, d_2, \ldots, d_n \rangle$
- Define *Coherence* as

$$Coherence(d_1, d_2, \ldots, d_n) = \min_{i=1}^{n-1} \sum_{w} \mathbf{1}(w \in d_i \cap d_{i+1})$$

Document Influence

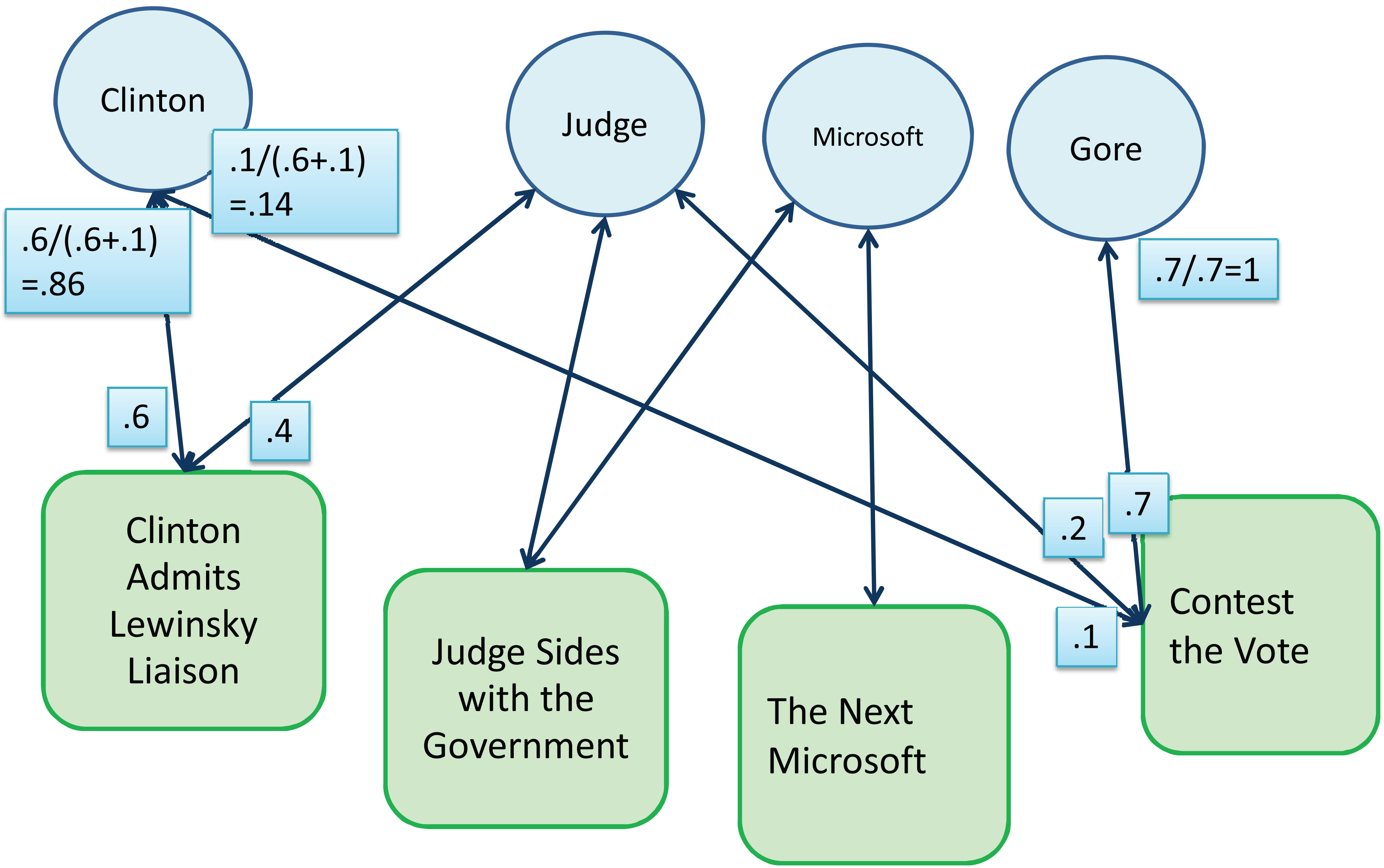
- The appearance of words is too coarse
  - Doesn't measure which words are important
    - Stop words are not important at all, other words can be very important
  - Important words might be missing from the articles
    - E.g. if the document has *lawyer* and *court*, also *judge* is probably important, even if it's not in the document
- The *influence* of d_i to d_{i+1} through word w is high if
  - d_i and d_{i+1} are highly connected
  - w is important for the connectivity

$$Coherence(d_1, d_2, \ldots, d_n) = \min_{i=1}^{n-1} \sum_{w} Influence(d_i, d_{i+1} \mid w)$$

Computing the Influence

- Measuring the influence is commonly done with linked data
  - E.g. PageRank computes an influence of the web page based on the link structure
- Here the news articles don't link to each other
  - The articles are joined via words in them
  - We want to assess the significance of a word for the link
- Build a bipartite graph of articles × words
  - Measure the influence of a word based on how surely we travel through it when moving from d_i to d_j
  - N.B. words can be influental even if they are in neither of the articles

Directed, Weighted Bipartite Graph



Shahaf & Guestrin 2010

Weights and Random Walks

- The document-to-word edge is weighted based on how important this word is to this document
  - E.g. TF-IDF
  - Weights are normalised so that each document's outgoing edge weights sum to 1
- The word-to-document edge uses same weights but normalised for words
- We consider random walks that start from d_i
  - If d_i is (strongly) connected to d_j, short random walks should visit d_j often
  - This probability is in the stationary distribution

Stationary Distributions

- The stationary distribution for random walks starting from d_i tells how big a proportion of time the walk stays in vertex v (an article or a word)

$$\Pi_i(v) = \varepsilon \cdot \mathbf{1}(v = d_i) + (1-\varepsilon) \sum_{(u,v)\in E} \Pi_i(u) \Pr(v \mid u)$$

  - ε is the restart parameter
    - we expect a re-start of the random walk after 1/ε steps
  - $\Pr(v \mid u)$ is the probability of moving from u to v
- We also compute the distribution with word w as a *sink*
  - $\Pr^w(v \mid u) = 0$ if $u = w$ and $v \neq w$, 1 if $u = v = w$, and $\Pr(v \mid u)$ otherwise

$$\Pi_i^w(v) = \varepsilon \cdot \mathbf{1}(v = d_i) + (1-\varepsilon) \sum_{(u,v)\in E} \Pi_i^w(u) \Pr^w(v \mid u)$$

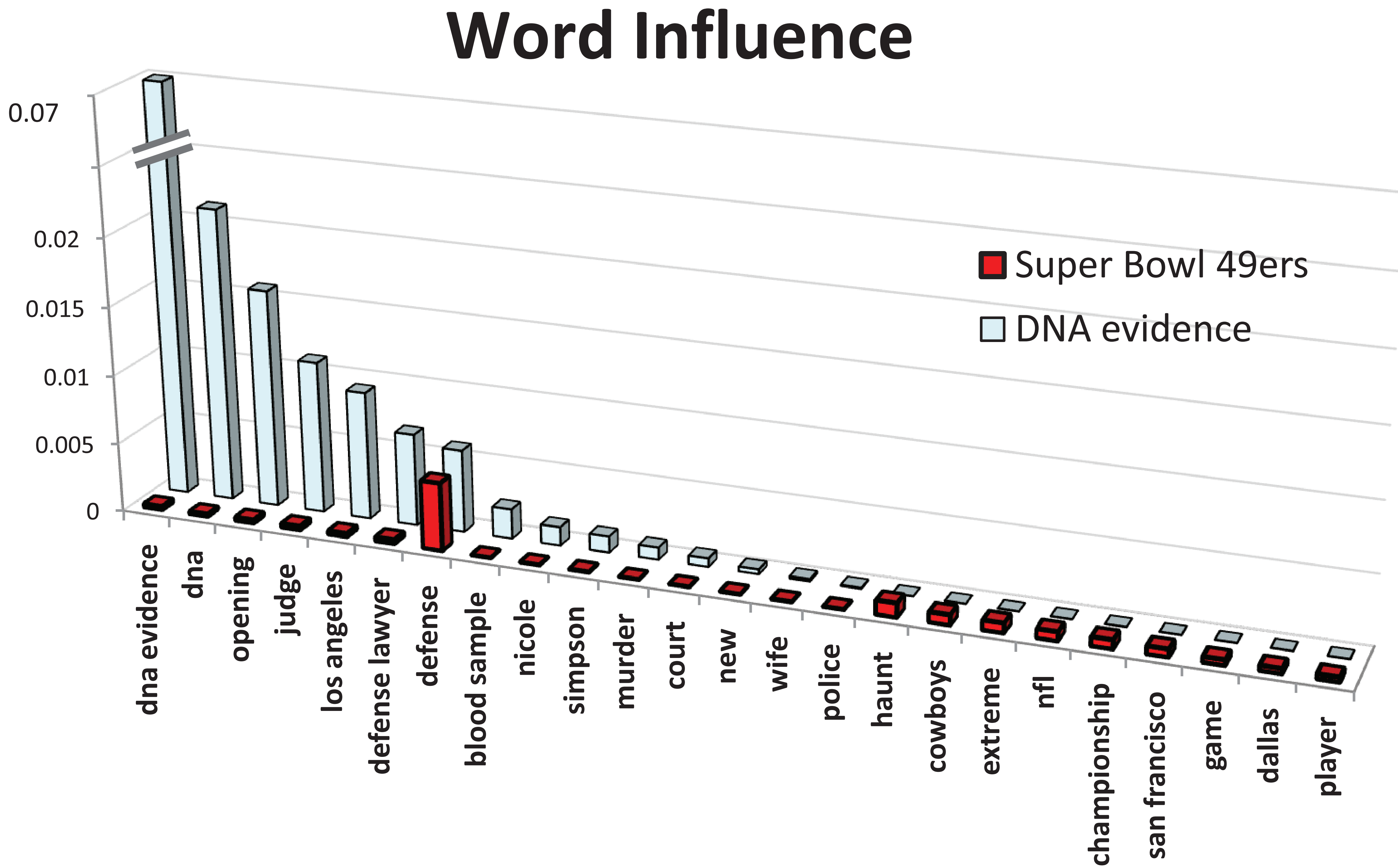
Computing the Influence

- We compute the influence as

$$Influence(d_i, d_j \mid w) = \Pi_i(d_j) - \Pi_i^w(d_j)$$

  - The fraction of time we spend in d_j if starting from d_i and walking thru w
- The stationary distributions can be solved using a power method
  - Start with uniform distribution, update the distribution, use that to update again, etc. until the updates converge
- The restart frequency ε matters a lot
  - Too small ⇒ too long walks ⇒ only general words matter
  - Too big ⇒ too short walks ⇒ only immediate words matter

Example



Influences of words on connections between an article about O.J. Simpson's trial and two other articles

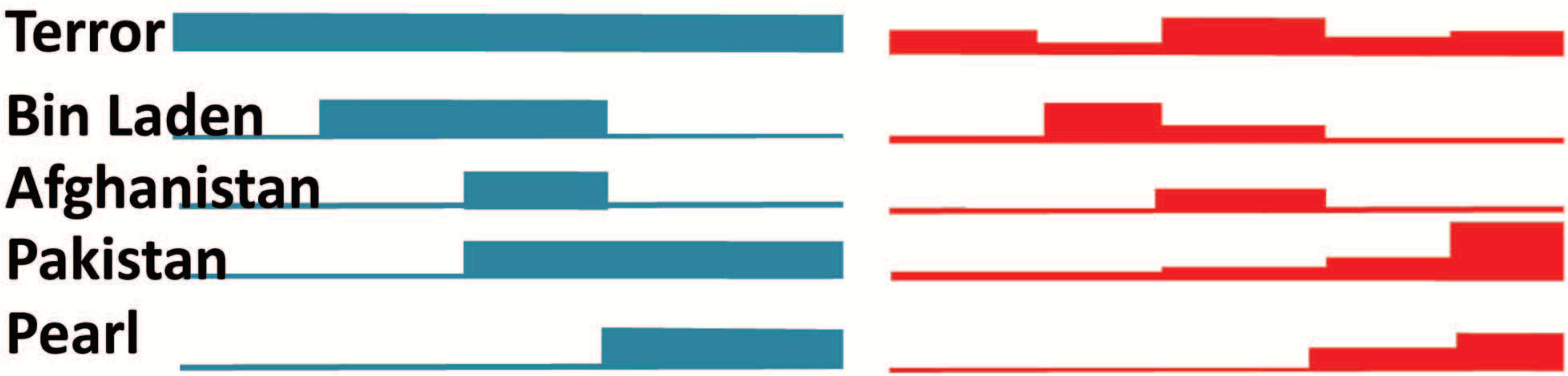
Shahaf & Guestrin 2010

Back to Coherence

- Recall, currently we define coherence as
  $Coherence(d_1, d_2, \ldots, d_n) = \min_{i=1}^{n-1} \sum_w Influence(d_i, d_{i+1} \mid w)$
  - This still suffers from *jitteriness*, jumping back-and-forth between topics
- We add the concept of *word activations*
  - Any word can be activated in any document
  - Each word can be activated only once
  - The total number of active words and the number of words active per transition is limited

$$Coherence(d_1, d_2, \ldots, d_n)$$
$$= \max_{\text{activations}} \min_{i=1}^{n-1} \sum_w Influence(d_i, d_{i+1} \mid w) \mathbf{1}(w \text{ active in } d_i, d_{i+1})$$

Activation Patterns Example



- Activation patterns connecting 9/11 to Daniel Pearl's murder
  - Left: activation patterns (documents on *x*-axis)
  - Right: activation patterns scaled with the influence
- "Terror" is constantly active
- There's a smooth chain between topics

Shahaf & Guestrin 2010

Scoring a Chain

- The optimal activation patterns for a given chain can be computed using an integer program
  - Includes the constraints for the activations
- But interger programs are NP-hard to compute
  - We can move to continuous activation levels (in [0,1]) to get a linear program
  - Now words can be activated multiple times
    - But only with fractional activation levels
- The number of active words in total (*kTotal*) and per transition (*kTrans*) effect the quality
  - Empirically $kTotal/4 \leq kTrans \leq kTotal/2$ is good

Finding the Chain: Idea

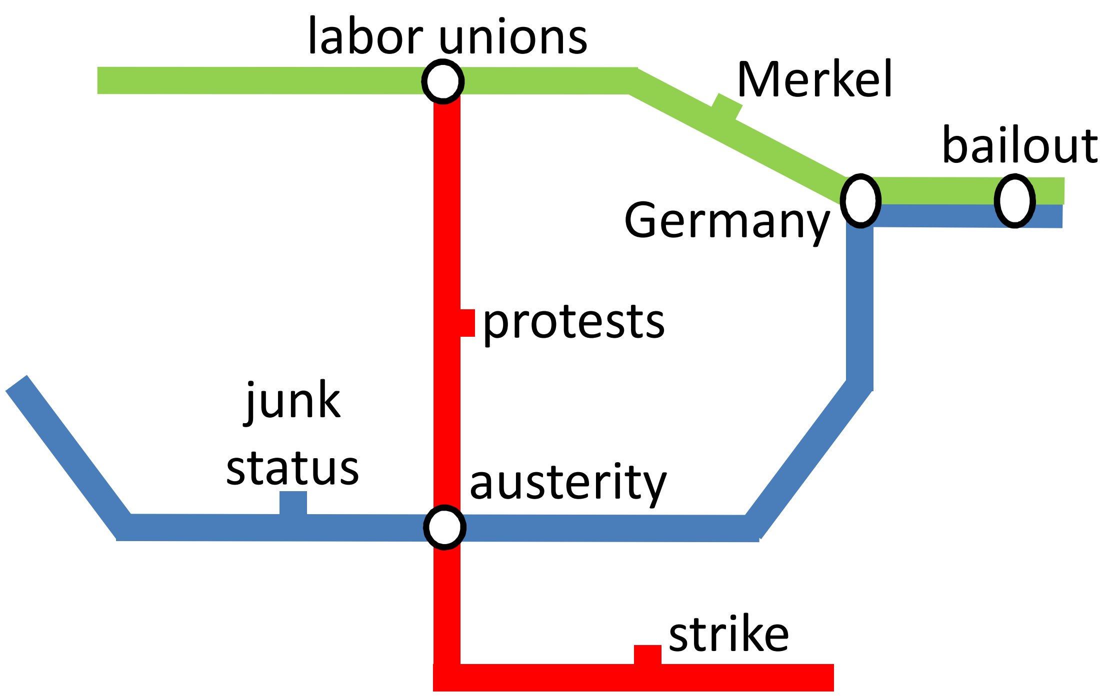
- We know how to score a given chain, but how to find one?
- Idea: find partial paths using optimistic approximations on their coherence
  - If p_i and p_{i+1} are two paths of length i and $i+1$ respectively and p_i is the prefix of p_{i+1}, then
  $$Coherence(p_i) \geq Coherence(p_{i+1})$$
  - If we extend p_i with edge e, the resulting path will have coherence at most
  $$\min\{Coherence(p_i), Coherence(e)\}$$
    - We only need to care about edges with high coherence

Finding the Chain: Algorithm

1. Compute all single-edge coherences and put the zero-edge path (*s*) to a priority queue *Q*
2. **while** *Q* is not empty
    2.1. Pop the highest-coherence prefix path from *Q*
    2.2. **if** path coherence has been approximated, compute exact and push the path back to *Q*
    2.3. **else**
        2.3.1. **if** this is *s*–*t* path, **return** it
        2.3.2. **else** compute all 1-extensions of the path that can reach *t* with remaining steps, approximate their coherence and push them to *Q*

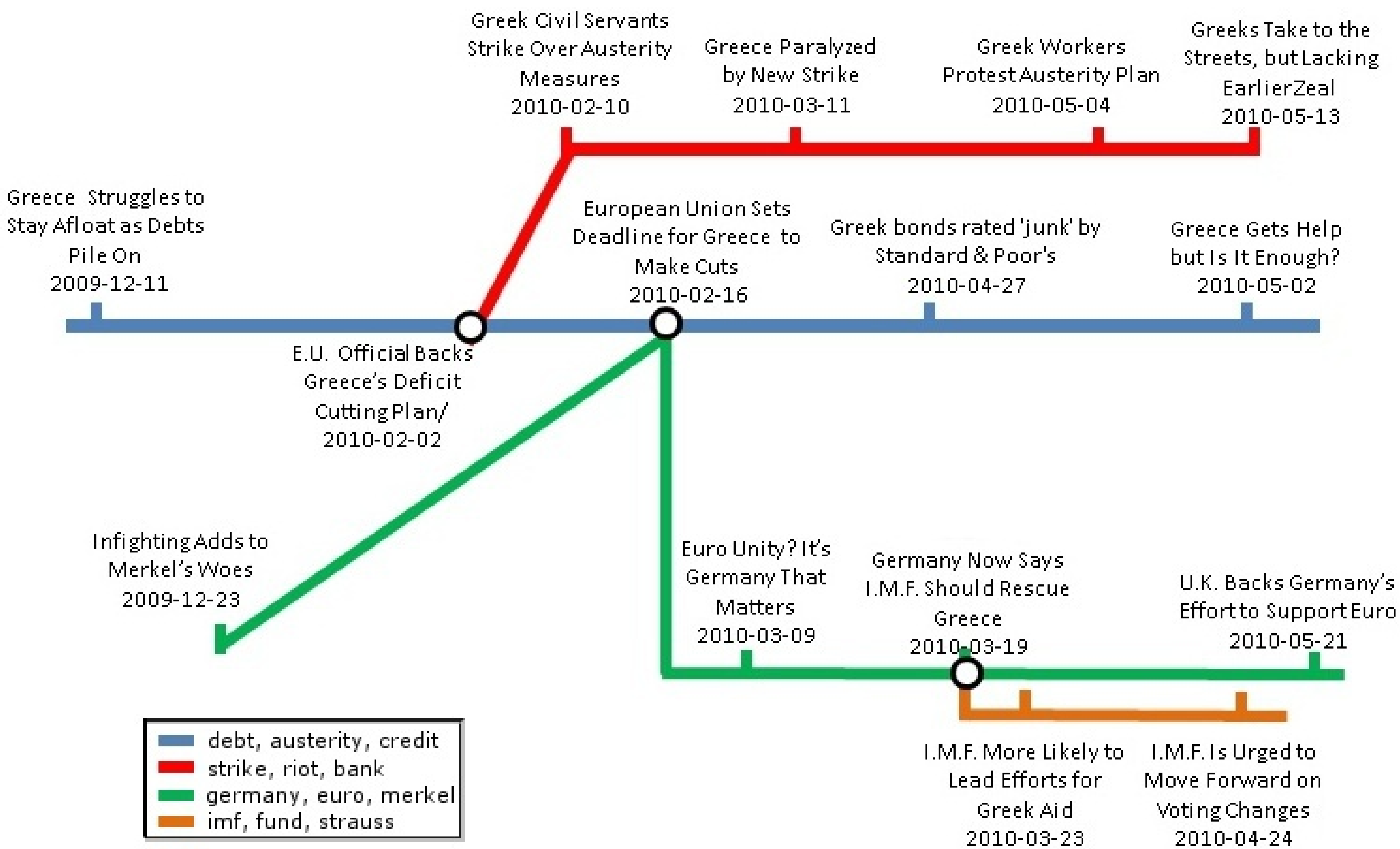
Metro Maps

- We've learned how to connect two news articles
  - But it still requires us to select those articles
- Could we map all connections within some topic?
  - Lines that explain progression of news (narrative)
  - Possibly intersecting and overlapping



Shahaf, Guestrin & Horvitz 2012a

More Detailed Example



Shahaf, Guestrin & Horvitz 2012a

Objectives for Metro Maps

- **Coherence**
  - Each line has to be coherent
- **Coverage**
  - Just asking for coherent lines yields very boring and narrow stories
  - We need the stories to cover many topics
    - Many stories and diverse stories
- **Connectivity**
  - The lines should connect to each other to reveal the structure

Coherence and Connectivity

- **Coherence** of each line is computed as when we were connecting the dots
  - Coherence of the map is the minimal coherence of any of its lines
  - We care about *m*-coherence: a line is *m*-coherent if each of it's sub-lines of length *m* is coherent
    - Makes computation simpler
- The **connectivity** of the map is the number of line pairs that intersect

Coverage

- Define $cover_d(w)$ be the amount document d covers word w (in [0,1])
  - E.g. a TF-IDF value
- The cover of a word w in map M is the probability that at least one document of M covers w

$$cover_M(w) = 1 - \prod_{d \in \mathrm{docs}(M)} (1 - cover_d(w))$$

  - Adding new documents that cover well-covered word doesn't help
- The **cover** of M is $Cover(M) = \sum_w \lambda_w cover_M(w)$
  - λ_w is a (subjective) word importance

Objective Function

- Coherence and coverage are constraints
  - We want lines to be coherent and have a good coverage, but we don't try to maximise either
  - Both have to be above some threshold
- We try to maximise connectivity within the given constraints
  - Coverage threshold stops us having just the same story many times
  - Coherence threshold stops us having meaningless crossings
    - Actually, *m*-coherence

Finding All *m*-Coherent Lines

- We generate all coherent lines of length m using similar best-first search as when connecting the dots
  - Priority queue of sub-chains, create all extensions of most-coherent sub-chain, remove chains of length m
- Of these we create a graph G
  - Each vertex is a coherent line of length m
  - There is an edge between two vertices if the corresponding lines differ in one document
    - The merge two such lines is still coherent
- This map gives us the input for our algorithm

Finding a High-Coverage Map

- From *G* we want to find a set of paths that maximise the coverage
- The coverage is *submodular* function
  - $f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$ if $X \subseteq Y$
    - "Diminishing returns"
  - We can get $(e - 1)/e$ approximation with greedy algorithm
    - But we cannot enumerate every candidate
- Compute the max-coverage path between every pair of documents and greedily select the best of them
  - Algorithms with $\alpha = O(\log OPT)$ approximation ratio exist
  - Overall, $(e^\alpha - 1)/e^\alpha$ approximation

Increasing Connectivity

- We now have coherent, high-coverage maps and we're left with maximising the connectivity
- We use local search
  - Replace each path of the map (one at time) with another one that increases the connectivity without hurting the coverage (too much)
  - After each replace has been tried, select the one with highest connectivity
  - Repeat until convergence
- Time complexity:
  - $|D|^m$ linear programs for coherence map creation
  - $K|D|^2$ quasi-polynomial algorithms for coverage
  - $K|D|^2$ quasi-polynomial algorithms for each iteration in local search

Essay Subjects for Topic II

- *Applications of frequent subgraph mining*
  - Read other literature; what is the data, how is it (modelled) as a graph, what are the subgraphs and why are they interesting
- *Metro Maps of Science*
  - Read *Metro Maps of Science* by Shahaf, Guestrin & Horvitz (KDD '12) and explain it
- *Parameters in Connecting the Dots and Trains of Thought*
  - Explain all user-supplied parameters in today's articles: what they do, why they are needed, how to find good values for them; give your opinion about these parameters (Too many/few? Easy/hard to understand the importance? etc.)

Feedback on Topic I Essays

- Good quality
- I could see your own ideas/opinions: good!
- Much improved citing practices
  - But: if you cite an article that has been published (in journal or conference), you have to give that information
    - And you don't have to give the URL where you found it (or access date)
  - It's important that the reader can understand what type of a work you're citing