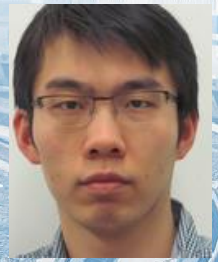
On Detecting Cherry-picked Trendlines

Abolfazl Asudeh, H. V. Jagadish, You (Will) Wu, Cong Yu

asudeh@uic.edu; jag@umich.edu; {wuyou, congyu} @google.com






Outline

- Bias in data presentation
- Problem Formulation
- Unconstrained Trendlines
- Constrained Trendlines
- Randomized Algorithms
- The most supported statements
- Experiments
- Discussions

Bias in Data Presentation

- To present the data in a way that it conveys a biased picture of a situation.

- Examples:

    - Bias through Visualization
        - Dishonest research
    - Bias through the choice of metric
        - Dishonest research
    - Bias through presentation ordering
        - Media
    - Cherry-picking data/facts
        - Politics

Fake news v.s. Biased data presentation:

- Fake news is a total fabrication → Can be detected/rejected by fact checking; “easy” to argue against
- Biased data presentation is based on “a grain of truth” (fact checking cannot reject it)
    - A popular techniques, especially in politics

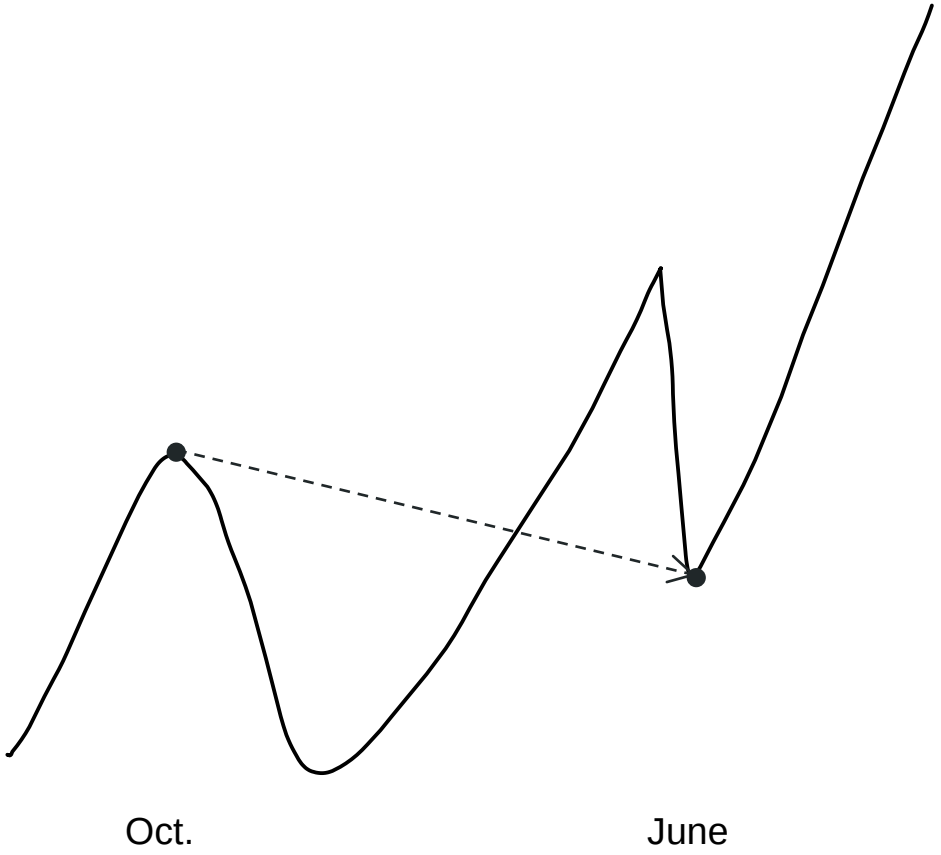
“A lie which is half a truth is ever the blackest of lies.” -- Alfred, Lord Tennyson

Cherry-picking Trendlines

- Politicians would like not to be caught blatantly lying, so they cherry-pick the factual basis for their conclusion.
- The points based on which a *statement* is made are carefully selected to show a misleading *“trendline”* that is not a *reasonable representation* of the situation.

Toy Example

- Unemployment rate has reduced by 5% from Oct. to June, which shows our policies have been successful in reducing the unemployment rate



Running Example

- It has been explained how cherry-picking short time-frames can distort the reality of global warming. The monthly climate data can be used to support the following fantasy-like claims:

- “summer was colder than winter in 2012 in the Northern Hemisphere” as, for example, the (average) temperature of Ann Arbor (MI, USA) on Aug. 18 (a summer day) was 58^F, whereas its temperature on Mar. 15 (a winter day) was 66^F.

Example 2: Giuliani's Adoption claim

- Rudy Giuliani claimed in the 2007 Republican presidential candidates' debate that ``adoptions went up 65 to 70 percent'' in New York City ``when he was the mayor''. The claim considered the total number of adoptions during 1996-2001 v.s. 1990-1995, while Giuliani was in office in 1994-2001.

Our goal is to *quantify* and *efficiently identify* such statements, made based on cherry-picked data

Discussions

- Cherry-picking has a long history and hence many different forms.
- In a nice article at PolitiFact, L. Jacobson goes over some of the examples of cherry-picking in US politics.
  - PolitiFact has reported cherry-picking ``***hundreds of times***'' in their fact-checks.
- While we believe our notion of support can be adopted for all cherry-picking settings, how to efficiently compute the support is problem-specific.
- For a large set of cases highlighted by PolitiFact, our algorithms can simply be adopted.

Problem Formulation

Data Model

- $x_1, \ldots, x_d$: trend attributes (e.g.: time, long., lat., city, etc.)
- y: Target variable (e.g.: adoption rate, temperature)

  - $p = < x_1, x_2 > = < July\ 20\ 2012,\ Ann\ arbor, MI >$
  - $y(p) = 70$
  - point $< \mathrm{p}, y(p) >$

Trendline

- we focus on trends derived by comparing a pair of points in data to make a statement.
- A trendline θ is a defined as a pair of trend points b (the beginning) and e (the end) and their target values in the form of

$$\theta = \langle\, (b, y(b)), (e, y(e)) \,\rangle$$

  - E.g., the trendline compares the temperature of Ann Arbor on two different days.

Trendlines based on Aggregate values

- made based on an aggregate over the target values in a window (with fixed/variable length)
    - e.g.: Giuliani's claim

- Fixed-size window length (e.g. 5-year terms): replace $y(x)$ with $\mathrm{Y}(x)$
- Variable-size window length: Add the window length as a new feature to x:
    - $p\langle x_1, x_2, \ldots, x_d, |w|\rangle, Y(p)$

Statement

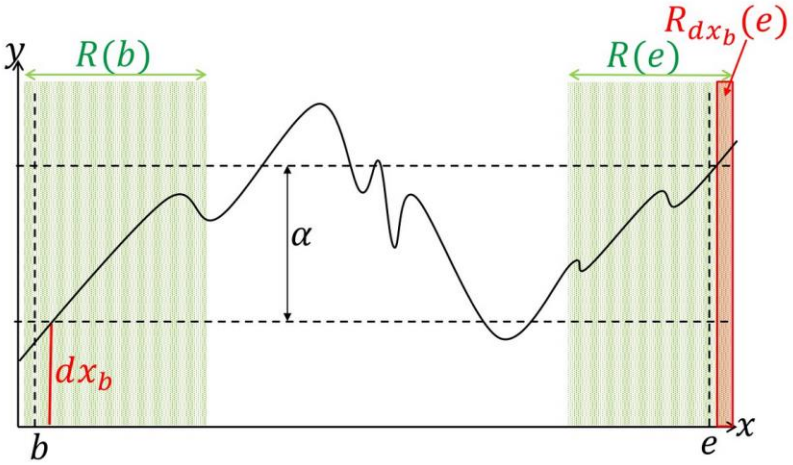
- Given a trendline θ, the statement S_θ is a range $S_\theta = (\bot, \top)$ such that
$$y(e) - y(b) \in (\bot, \top)$$

- In the running example:
  - b =Aug. 18 2012 (summer), Ann Arbor – MI, $\mathrm{y(b)} = 58$^F
  - e =March 15 2012 (winter), Ann Arbor - MI, $\mathrm{y(b)} = 66$^F
  - Statement: summer was colder than winter, is: $S_\theta = (0, \infty)$ which is satisfied by θ since
$$y(e) - y(b) = 66 - 58 > 0$$

Support Model

- Observation: if a statement is not based on a cherrypicked trendline, other data points should also ***support*** it.
  - cherry-picked trendlines are carefully selected and, therefore, may change by slightly changing the trend points.
  - In the running example, perturbing the beginning and/or the end points of the chosen dates by even a few days results in trendlines that do not support the statement.

Support Model



- Support Region: a neighborhood around the selected trend points
- Support of a statement: The ratio of the “valid” trendlines in the support region for which their target value difference remains within the acceptable range.

$$\omega(S, R_S, \mathcal{D}) = \frac{vol(\{valid\ \langle p \in R(b), p' \in R(e)\rangle \mid y(p') - y(p) \in (\bot, \top)\})}{vol(\{valid\ \langle p, p'\rangle \mid p \in R(b), p' \in R(e)\})}$$

Problem Formulations

1. Compute the support of a statement

2.

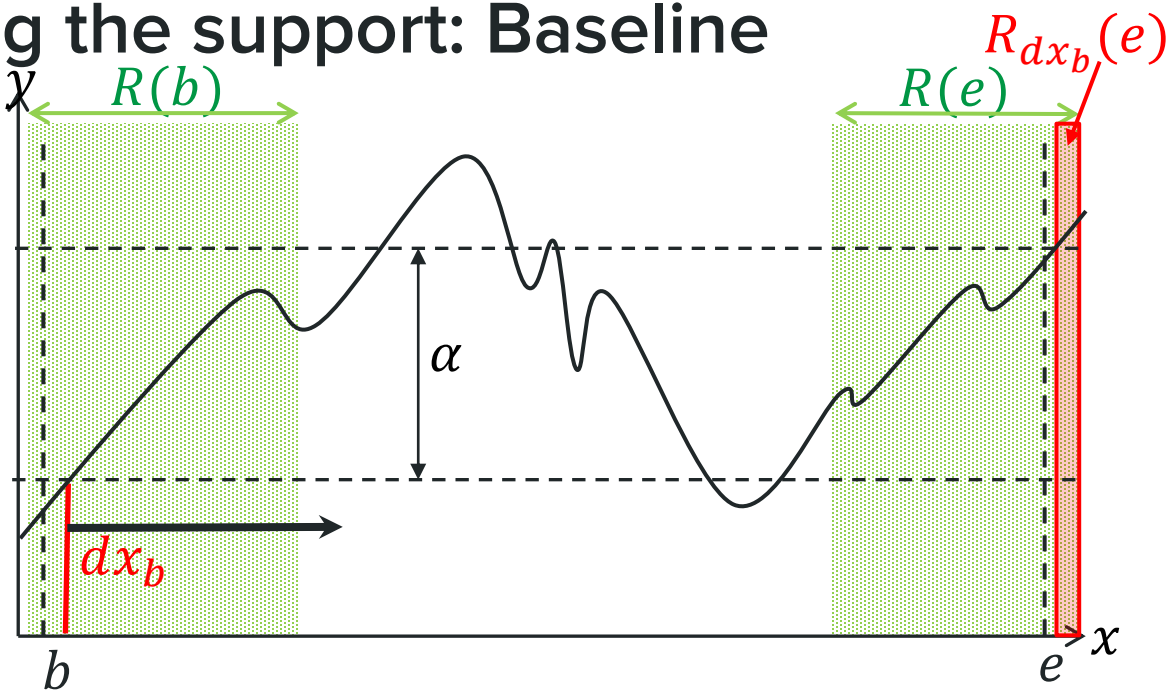
   Find the most supported statement for a given range

   Find the tightest statement for a given support value

Unconstrained Trendlines

Computing the support: Baseline

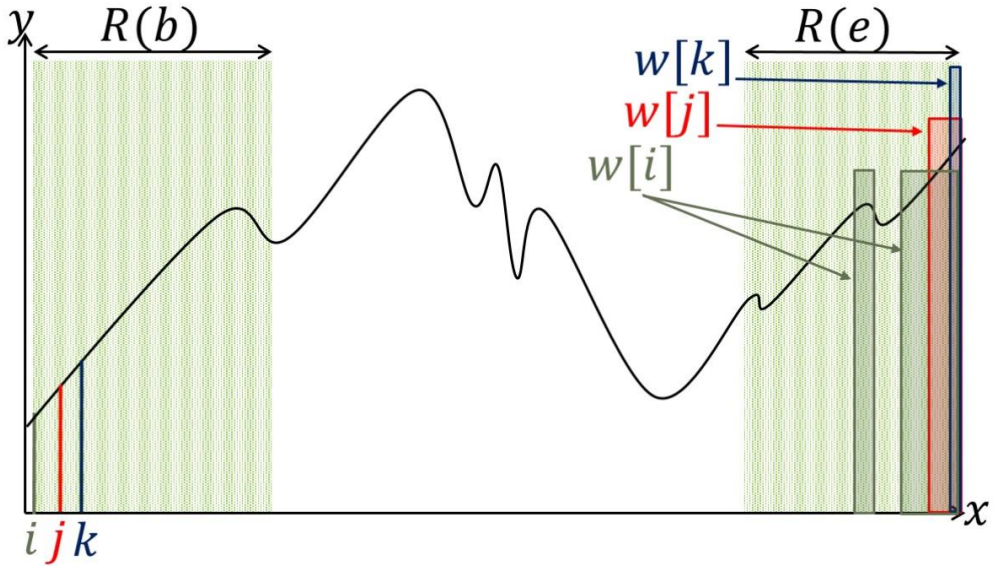
$O(n^2)$



Efficient Algorithm

- For every point $d_i \in b$, define w_i as the number of points in b for which $y(d_i) - y(d_j) \in (\perp, \top)$. Then support of a statement can be computed as

$$\sum_{\forall d_i \in b} w[i]$$



Efficient Algorithm

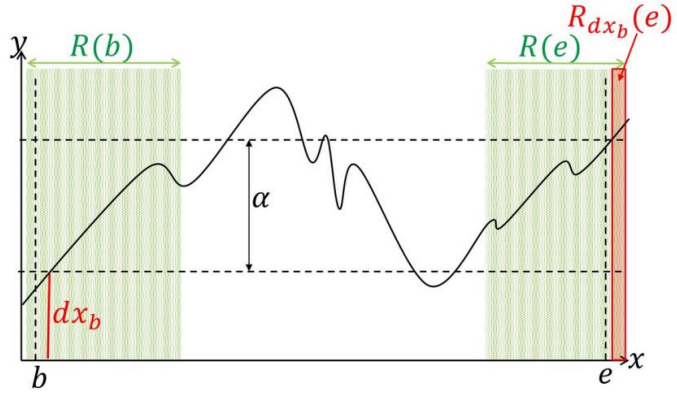
O(n logn)

- Design the cumulative function $F(y) = |\{dx \in R(e) \mid y(dx) < y\}|$

Sort O(n log n)

- Using *F*, $w[i] = F\big(y(dx[i]) + \top\big) - F\big(y(dx[i]) + \bot\big)$

Binary search O(log n)



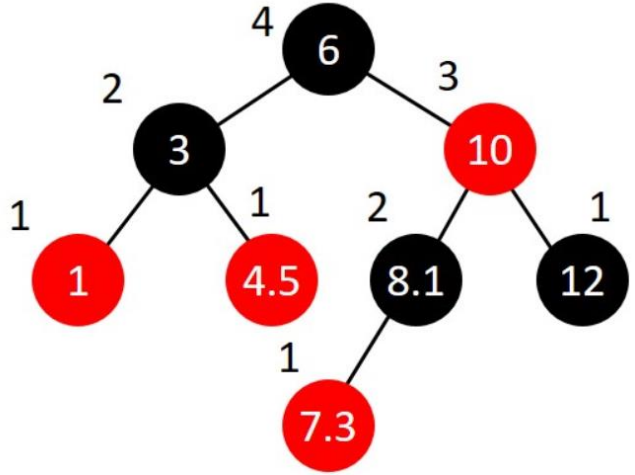
Constrained Trendlines

- Unconstrained trendlines: every pair of points in the beginning and end range are *valid*
- Constrained trendlines: The choice of the point in the beginning range, limits the end-points
    - Extreme case (Single-point enforcement): trendlines should have fixed length. The only valid end-point is $\mathrm{e} = b + l$
    - General case: The length of trendlines is limited to a range of values

Challenge

- The choice of the beginning limits the choices in the end → the cumulative function method does not work any more

- If window size < log(n): Baseline is already in O(n log(n))
- If not?



Randomized Algorithms

Large –scale settings

- How can we quickly estimate the support value of a trend?

- Monte-Carlo methods

    1. Generate unbiased samples
    2. compute the ratio of cases that satisfy the requirements (support the statement)

Randomized Algorithms

1. Pair Sampling

2. Point Sampling ← practical solution

Mining Problems

- Question: If the given statement is not “fair”, what is a fair statement to make?

- Example: In the running example, what are the fair statements about the temperature?

- Most Supported Statement

Given the width of the support range, what is the statement that has the maximum support

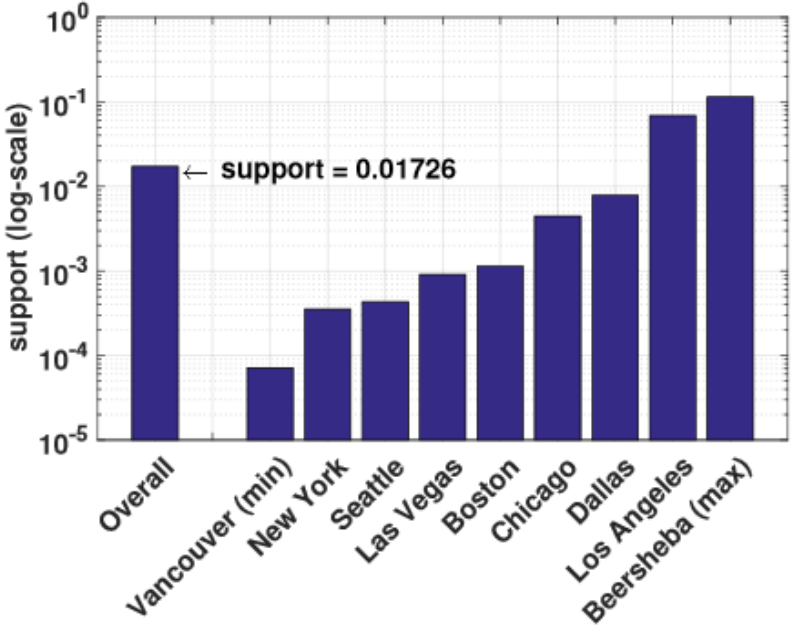
- Tightest Statement

Give a support value (e.g.: 80%), what is the tightest statement with the given support.

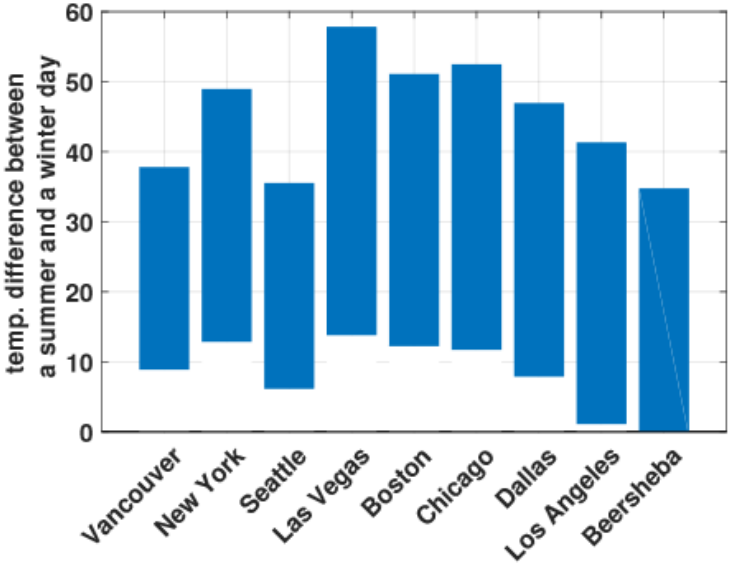
Experiments

Experiments, Proof of Concept (running example)

Support of (winter colder than summer)

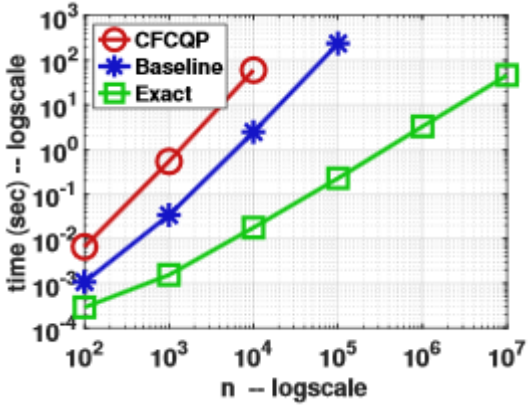


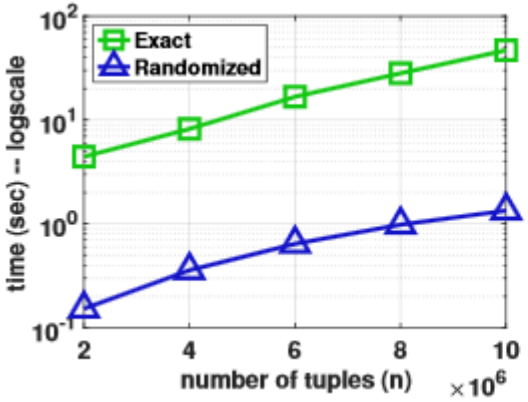
Tightest Statement with support 0.8



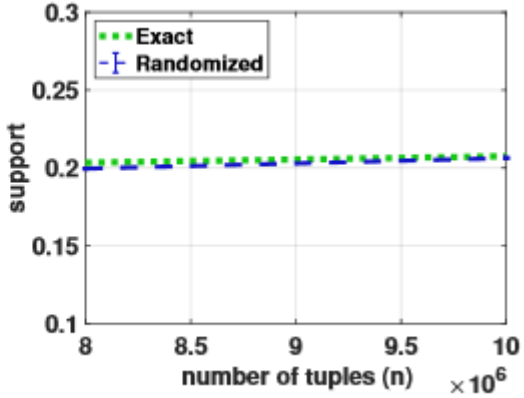
Performance Evaluation

Exact





Randomized



Thank you

- Abolfazl Asudeh, asudeh@uic.edu, www.cs.uic.edu/~asudeh/,  @ab_asudeh
- H. V. Jagadish, jag@umich.edu, web.eecs.umich.edu/~jag/
- You (Will) Wu, wuyou@google.com , research.google.com/people/YouWillWu/
- Cong Yu, congyu@google.com , sites.google.com/site/congyu/

Discussions