A Meta-Analysis of Metrics for Change Point Detection Algorithms

Master's Thesis Defence - Matt Chapman - August 2017

Presentation Abstract

- There is little consensus on the “best” way to measure accuracy of change point detection algorithms
- Simulation studies show that popular measures can be ineffective in some situations
- Change point analysis of real-world data shows disagreements between metrics
- Effectiveness of algorithms on real-world data used in this project is questionable

Project Motivation

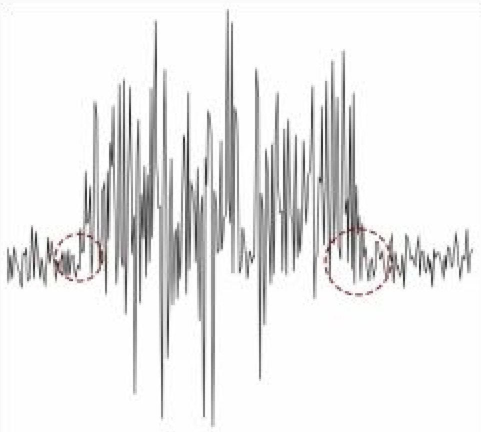
- Buzzcapture BV - the project host:
    - Online reputation specialists
    - Main software product is a dashboard for monitoring online and offline media
- The problem:
    - How do you detect when a conversation “goes viral”?
- A possible solution:
    - **Change point detection algorithms**
    - Change point analysis also incorporates anomaly/outlier detection, edge detection, and various other synonyms.

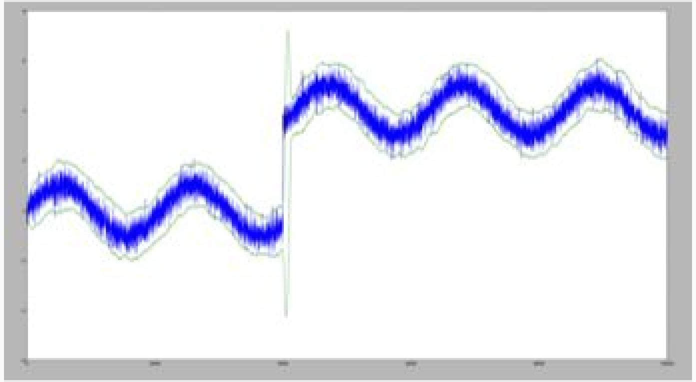
Change Point Analysis

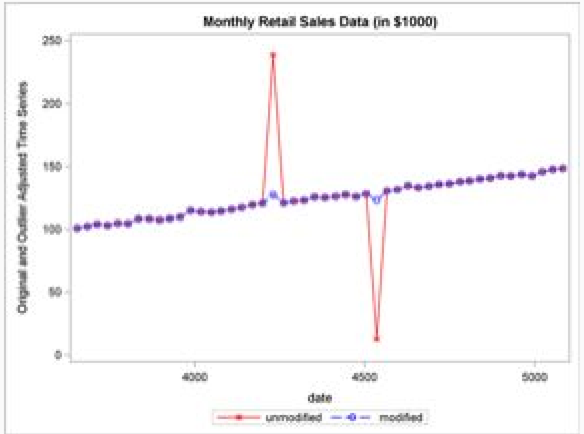
Fairly self explanatory:

Involves detecting some "change" in the properties of data.

For example, a change in mean, variance, correlation, or any other measurable statistic.







Change Point Analysis

- First paper on the subject was in 1954, by E.S. Page
- Titled “Continuous inspection Cycles”
- Heavily motivated by quality control in manufacturing
- Field has now grown to encompass many fields:
  - Linguistics
  - Intrusion Detection
  - Spam Filtering
  - Medical Diagnostics
- Over 565 publications on the subject[1]

This then posed a problem with the original project plan:

- Intention was to create a method for change point detection in Buzzcapture’s data
- This turned out to not have much value as a research project
- Another approach was needed

1. Killick, R., Nam, C. F.H., Aston, J.A.D. and Eckley I.A. changepoint.info: The changepoint repository (2012) http://changepoint.info

Pivoting the Project

If creating a method for change point detection wouldn’t be novel or interesting, what would be?

New thoughts then came about:

- How would you prove a method is better than another? **Metrics**!
- What metrics are utilised?
- Are these metrics actually useful?
- If we can break one of these metrics, *that* would be novel!

Research Question:
Are existing metrics in the field of change point detection effective and accurate?

Derived Research Sub-Questions

1. In what way are existing metrics **deficient** when applied to change point detection problems?
2. Do existing metrics agree on the **best approach** when used to evaluate change point detection algorithms applied to real-world data?
3. Is there a metric **more suited** than the others, for the purpose of evaluating change point detections according to functional requirements set forth by the project host?

Derived Research Sub-Questions

4. What would an **ideal metric** for evaluating change point detection approaches look like?
5. Do metrics show that change point detection is a **reasonable** and **effective** approach for the use-case of the host organisation?

Getting to work...

Which algorithms should be used?

- Pruned Exact Linear Time (Killick, Fearnhead & Eckley, 2012)
- Segment Neighbourhoods (Auger & Lawrence, 1989)
- Binary Segmentation (Jackson et. al., 2005)

All widely used, all well documented, and most importantly, all with reference implementations available in **R**.

Choosing Evaluation Measures

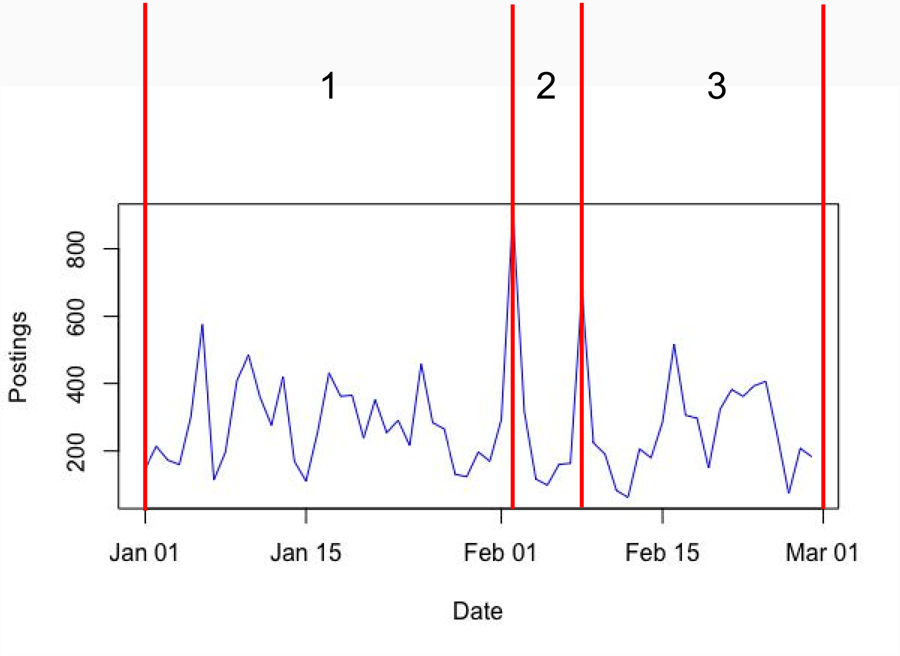
A literature study showed the following:

- Many publications favoured binary classification approaches:
  - Calculating Recall, Precision, F-Score
  - Receiver Operating Characteristic curves - plotting false/true detection rates
- At least one publication utilised **clustering measures**:
  - Rand Index
  - Adjusted Rand Index

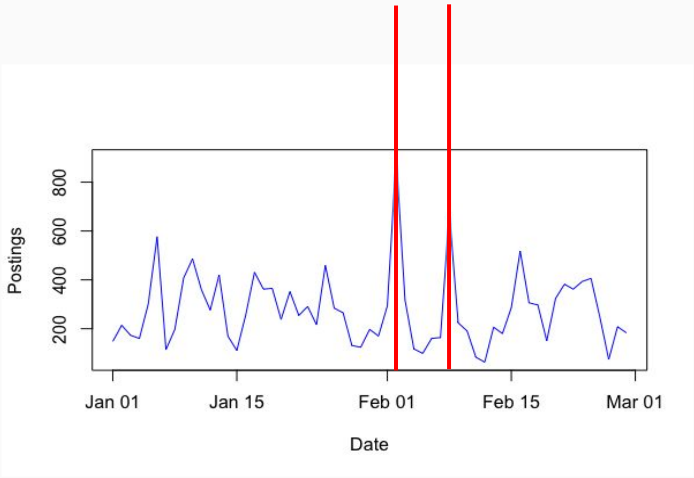
Clustering Measures

- This was novel - not many publications using this for evaluation of change point problems
- This opened an opportunity to see if a different clustering measure would perform well in this domain:
    - BCubed
    - Designed by Bagga & Baldwin in 1998
    - Computes “correctness”, “precision” & “recall”, before taking a harmonic mean in the same way as the binary classification “F-Score”

Calculating Clustering Measures by Segmentation



Binary Classification in Change Point Detection



Correct

False Negative

False Positive

Label	0	0	0	0	0	1	0	0	1	0	0	0
Classifier	0	0	0	0	0	1	0	0	0	1	0	0

1 = change point

0 = no change point

Final listing of algorithms & metrics

Change Point Detection Algorithms:

- Pruned Exact Linear Time
- Segment Neighbourhoods
- Binary Segmentation

Using the following test statistics

- Mean
- Variance
- Mean & Variance

Evaluation Metrics:

- Binary Classification:
  - Precision
  - Recall
  - F-Score
- Clustering:
  - Rand Index
  - Adjusted Rand Index
  - BCubed

Algorithm Configuration

Change point detection algorithms require the following:

- Penalty Function - to “optimise” the number of detected change points
  - Schwarz Information Criterion chosen for these studies
- Assumed distribution of data
  - A “normal” distribution is assumed for these studies

Designing the Experiments

Simulation Studies

- Designed to decouple the metrics from the algorithm implementations
- A simple data set and a result from a “pseudo algorithm” are simulated in various ways, and the metrics calculated at various points in the simulation
- This way, we can see what, if anything, causes the metrics to “break”.

Properties Tested by Simulation Studies

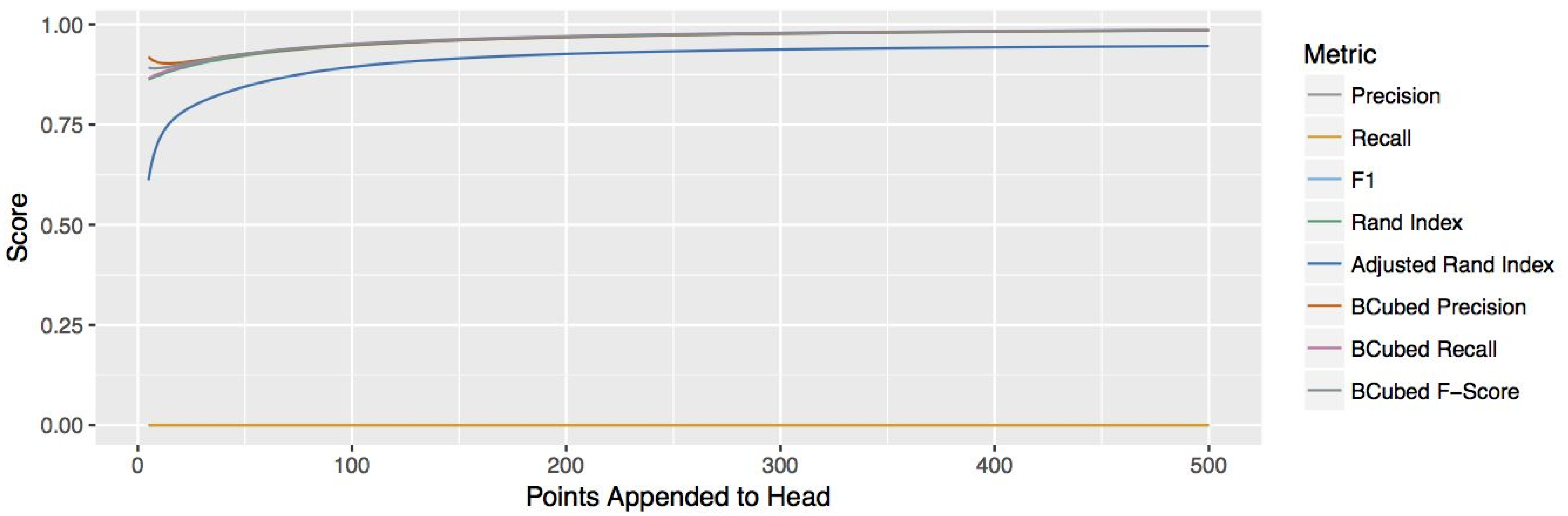
- Dependence on sample size
- Impact of data preceding/following a known change point & its detection
- Ability to apply a temporal penalty for early/late detections
- Ability to penalise for false positives
- Ability to penalise for false negatives
- Impact of change point density

Simulation Studies

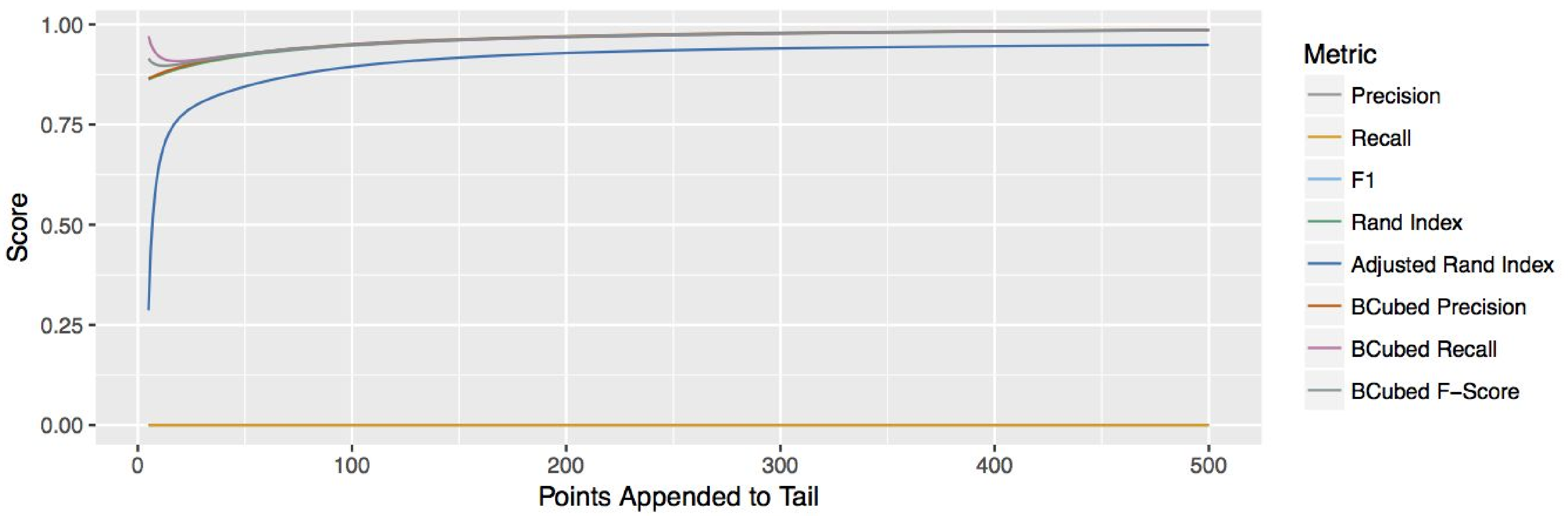
1. Increasing the “head” of the data
2. Increasing the “tail” of the data
3. Moving the “true” **and** detected change point through the data
4. Moving a detected change point through the data (temporal penalty)
5. Adding false positive detections
6. Adding false negative detections
7. Varying change point density in fixed-length data
8. Varying change point density in variable-length data

Simulation Study Results

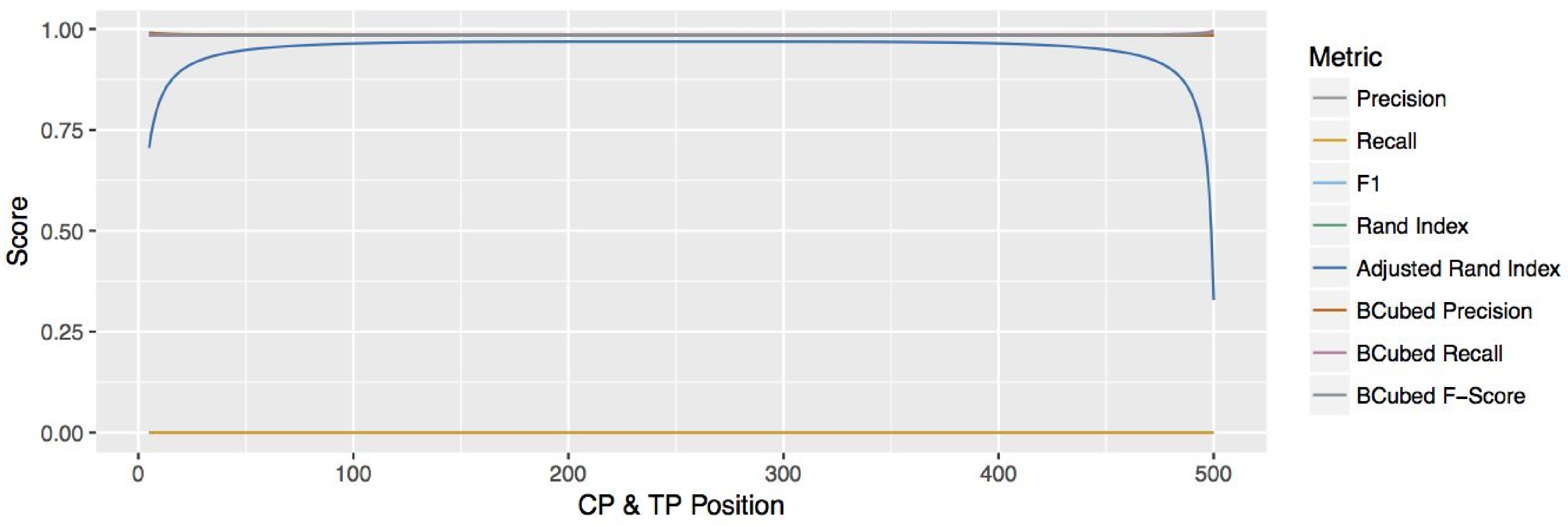
Adding Points to the “Head”



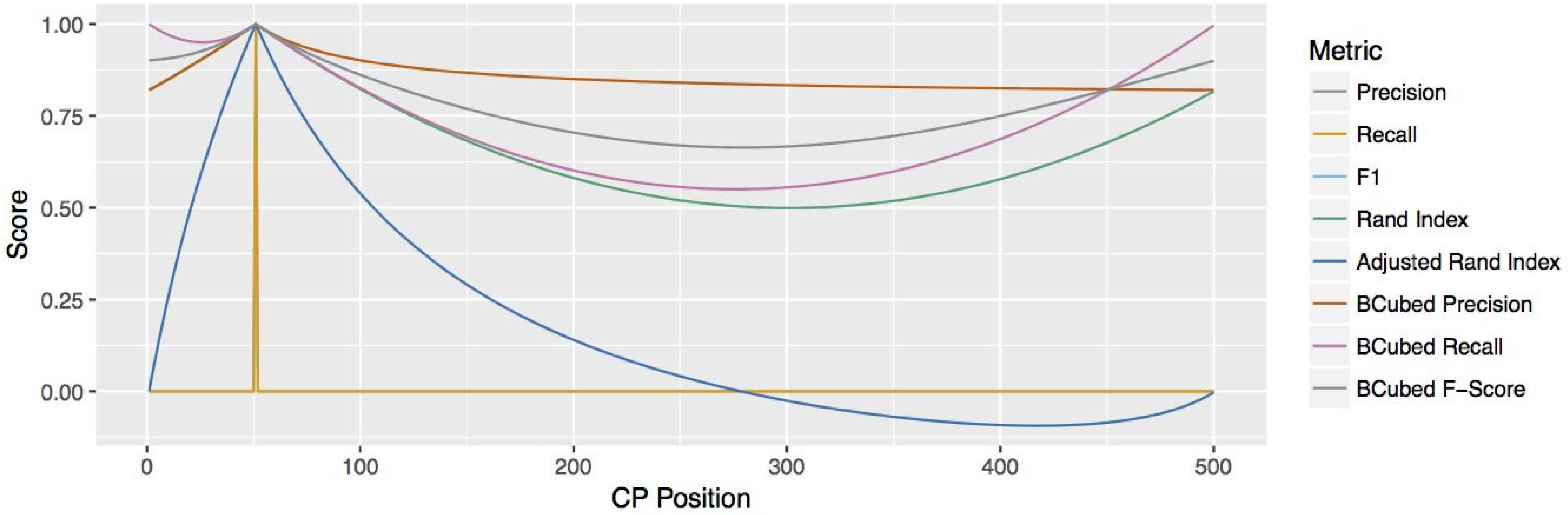
Adding Points to the “Tail”



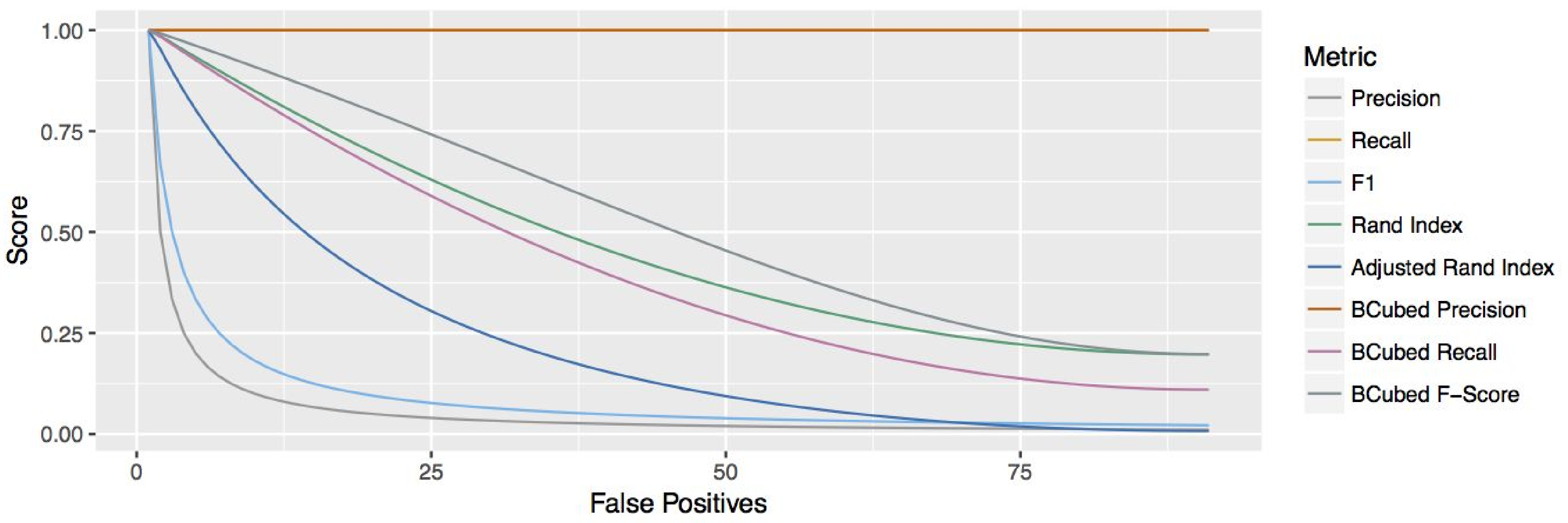
Moving the True Change Point and Detected Change Point



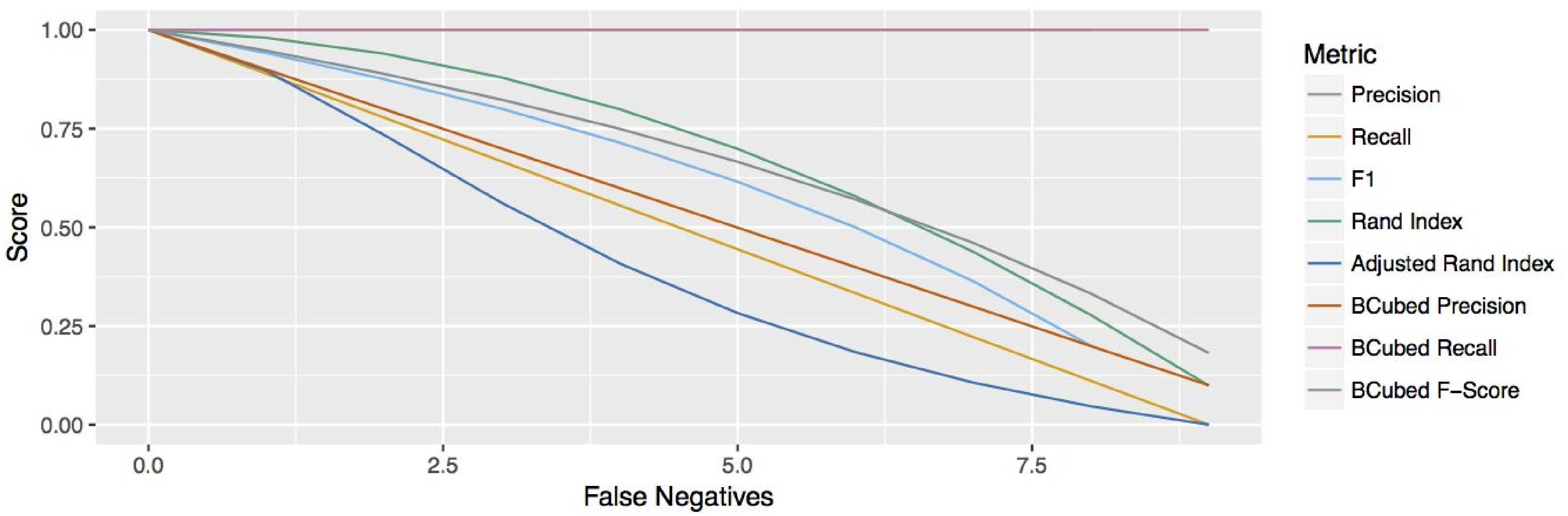
Applying a temporal penalty



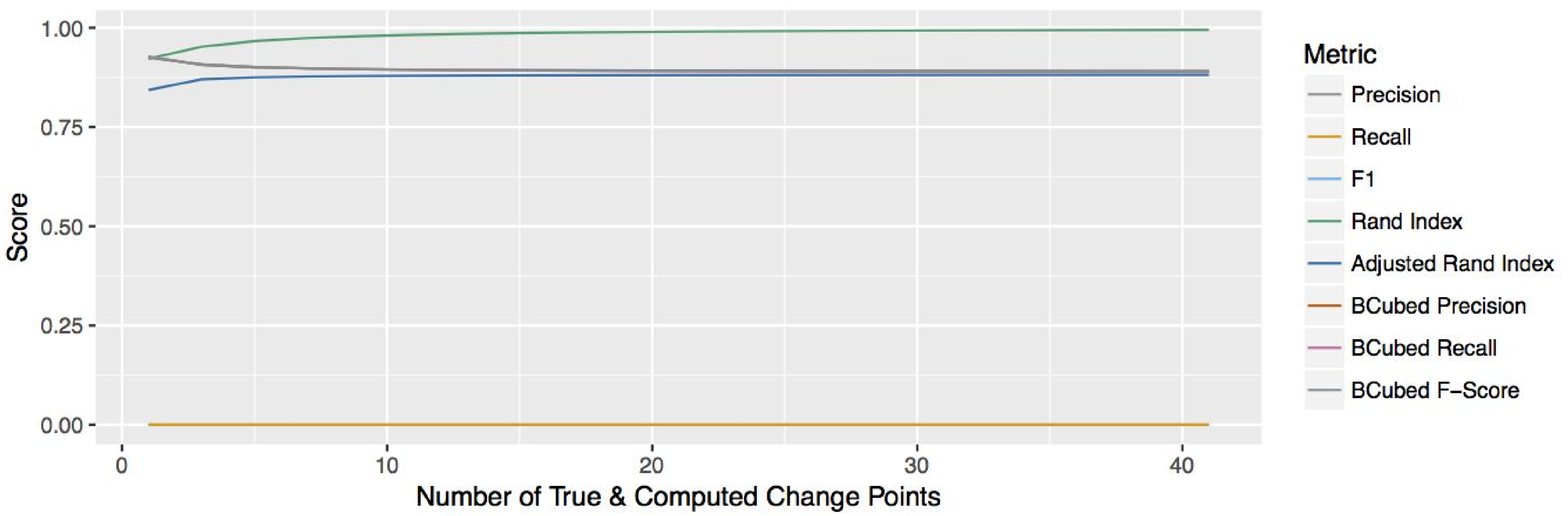
Adding False Positives



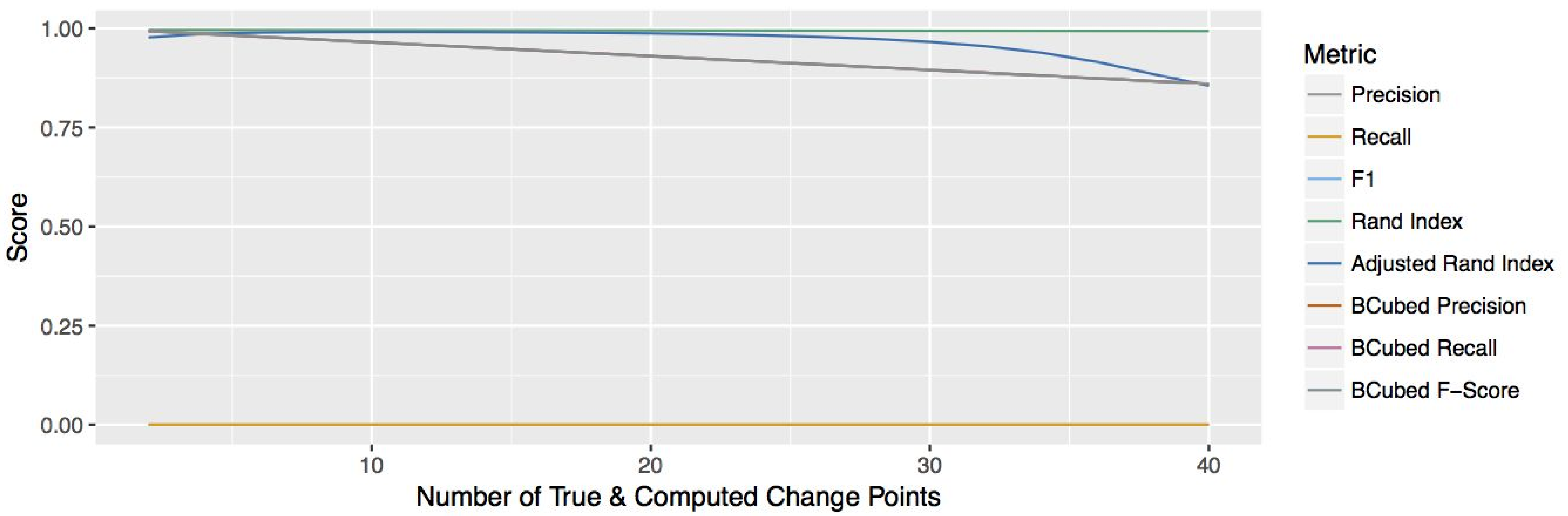
Adding False Negatives



Change Point Density (Variable Length)



Change Point Density (Fixed Length)



Conclusions from Simulation Studies

- All tested metrics behave “badly” in some situations:
  - Data set length affects scoring
  - Change point “density” affects scoring
  - Strange behaviour exhibited by measures WRT temporal penalty application - metric value *increases* as the detection becomes later!
- Thus, I conclude that none of the metrics in this study are ideal or maximally effective for evaluating change point algorithm performance.

Real-World Data Analysis

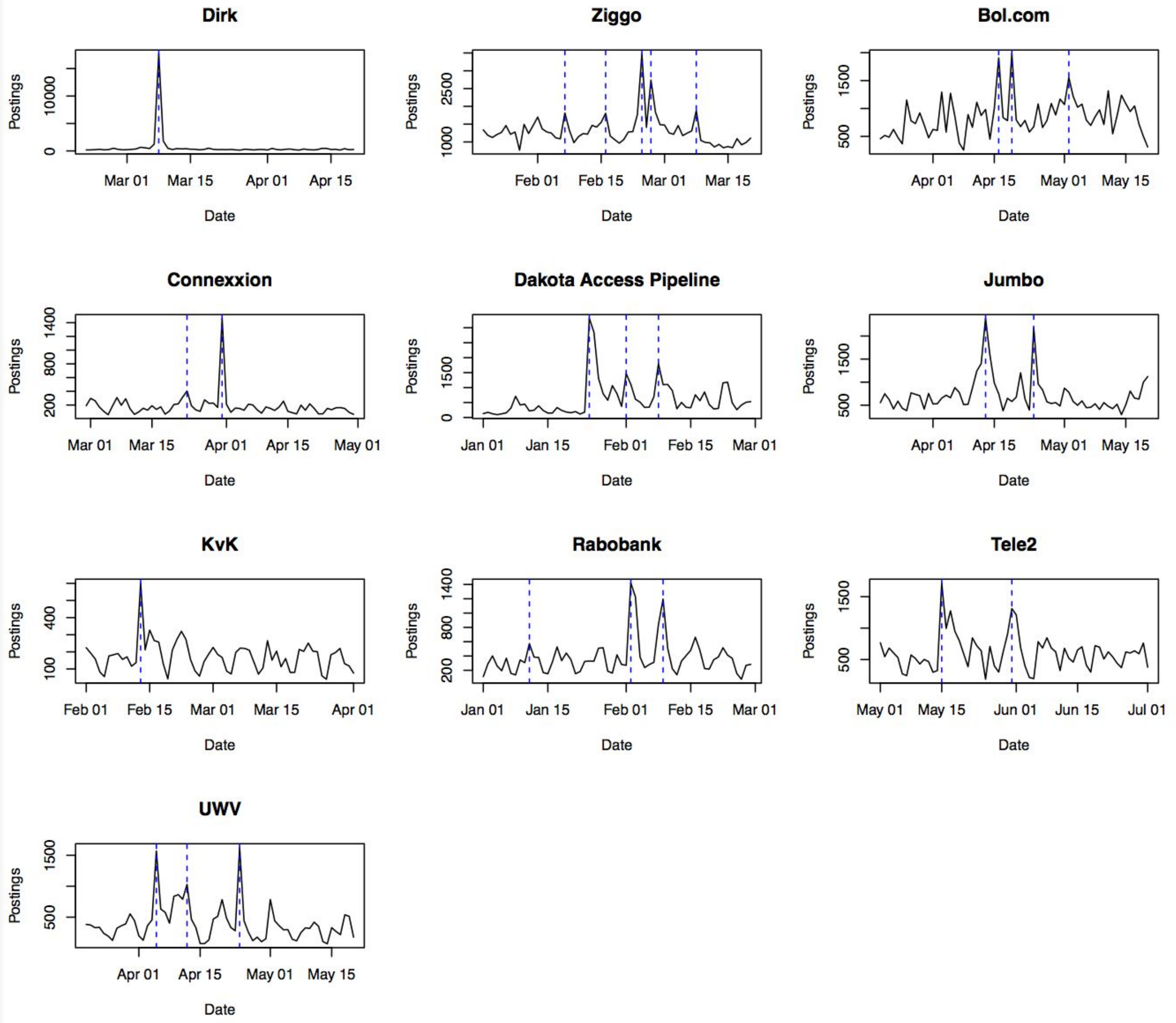
Designing the Experiment

- Analysts employed at Buzzcapture were asked to provide examples of clients that have had “significant” changes in conversation volume/postings volume recently
- This request resulted in 10 different data sets corresponding to various brand names, being selected
- Buzzcapture’s Head of Research then annotated the data-sets with where they expected changes to be detected - providing the ground truth for this study.

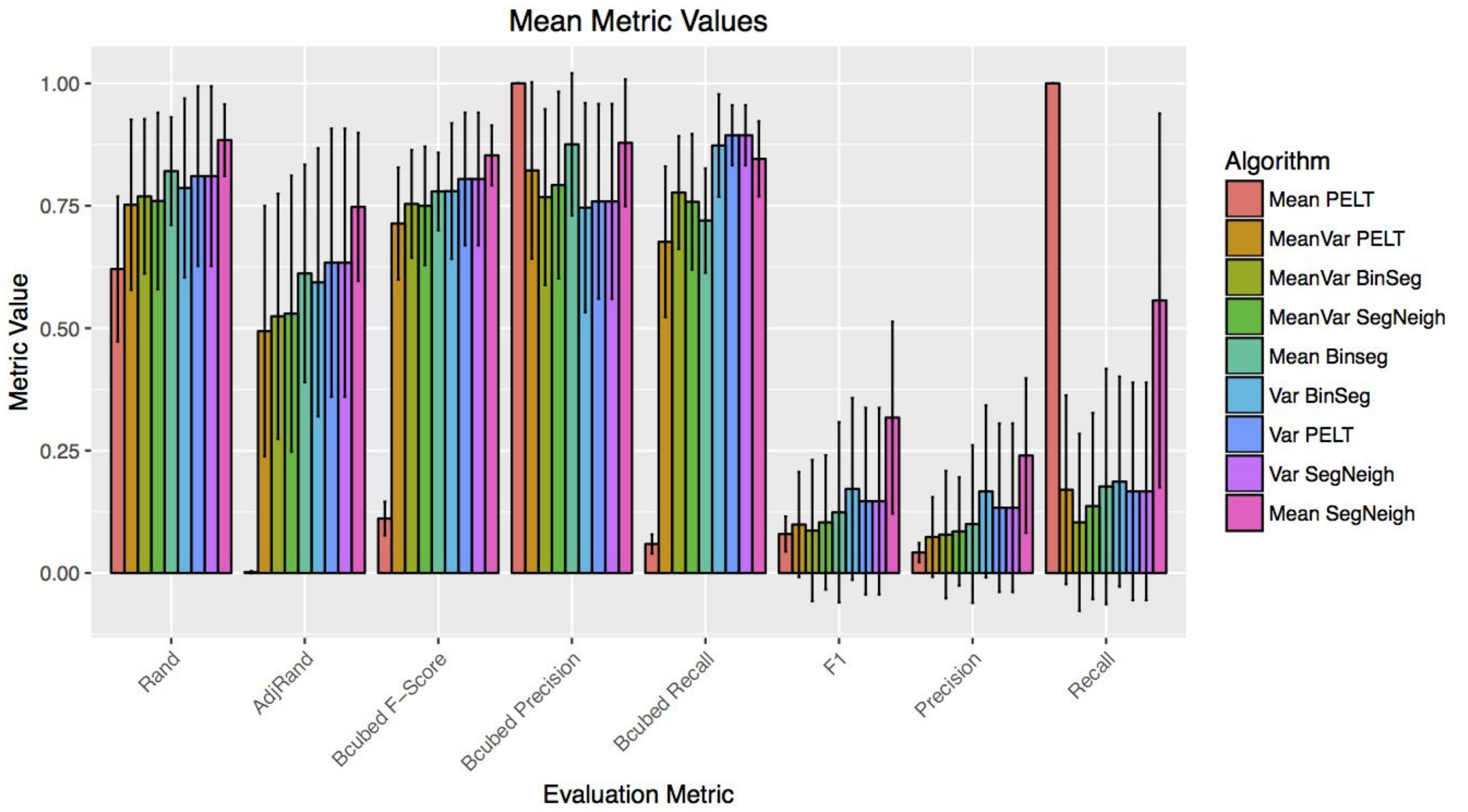
Designing the Experiment

- Data sets cover 60 days of conversation/posting volume for each brand
- Chosen to provide a good mix of large/small spikes in activity, noise, etc.
- Each algorithm is run **3** times against each data set testing for changes in:
  - Mean
  - Variance
  - Mean & Variance
- For each run, all of the aforementioned metrics are calculated
- Finally, the mean of each metric is taken for every data set, and 1 standard deviation is calculated

Data Sets & Ground Truth Annotations



Real World Data Results



Real World Data Results

- At first glance, the metrics appear to agree with each-other on the **best** algorithm/test statistic.
- But, do they really? To test this, all algorithms ranked according to each metric
- Calculating Kendall's Tau for every pair of metrics suggests some disagreements:
  - The only agreements are
    - Rand Index & Adjusted Rand Index
    - F1 Score & BCubed F-Score
  - Nothing else agrees with $p < 0.05$

Real World Data Conclusions

- The algorithms didn’t perform very well with this data
  - This doesn’t mean the algorithms don’t work - just that more work needs to be done with penalty values and distribution assumptions to find “most effective” method
- The metrics disagree with each other on which algorithm is “best”

Answering the Research Questions

In what way are existing metrics deficient?

- Clustering measures are significantly affected by data set size & change point density
- Clustering measures exhibit inconsistent behaviour when penalising for late or early detections
- Binary Classification metrics performed well in simulation studies, but were not so effective for measuring performance in real world data study.

Do existing metrics agree on the “best” approach for our data?

Yes & No.

- There is an agreement on the “top ranked” approach, but lack of correlation between rankings bring this into question.
- There were situations where algorithms performed badly when results examined by-eye, but performed well according to metrics.

Was 1 metric better than the others, given a set of requirements?

- None of the metrics were appropriate for evaluating all criteria
- This, combined with the poor performance of change point detection methods with this data, suggests that change point detection wasn't the best approach for useful notifications regarding impending virality of a given conversation.

What would an ideal metric look like?

- Credit for correct detections, penalise for incorrect detections
- Large penalisation for missing “relevant” changes, small penalisation for missing “less relevant” changes
- Unaffected by data set size
- Unaffected by change point density
- Plot of “score” against distance from true change point should be linear

Is change point detection an effective approach for this use-case?

Inconclusive.

- Some situations in which the algorithms performed well
- Conversely, also some situations in which they did not
- Conclusion brings more questions to the table:
    - Did algorithm configuration have a large impact on the results?
    - Was the penalty term selection sub-optimal for this domain?

Final Conclusions

Final Conclusions

- This thesis focussed on evaluating change point detection algorithms.
- Simulation studies showed some metrics “misbehave” in certain situations
- Real-world data study showed metrics **disagree** on ranking algorithms
- Real-world data study also showed mixed results for effectiveness of algorithms
- More in-depth study required (with larger sample sizes) to conclude with more confidence that change point detection is not effective for this use-case
- Further studies to cover (many) more simulations would be useful

Questions?