Temporal Graphs

KRISHNAN PANAMALAI MURALI
METRICFORENSICS: A Multi-Level Approach for Mining Volatile Graphs

Authors: Henderson, Eliassi-Rad, Faloutsos, Akoglu, Li, Maruhashi, Prakash and Tong.

Published: KDD 2010

METRICFORENSICS is a scalable framework for analyzing volatile graphs.

Volatile graphs are a stream of duration stamped edges in the form:

$\langle v_{src}, v_{dst}, \text{start\_time}, \text{duration} \rangle$

Examples are physical proximity graphs and IP-to-IP communication graphs.
METRICFORENSICS: Key contributions

• METRICFORENSICS provides:
  • A multi-level ‘drill-down approach’
  • Collection of analysis techniques
  • User-selected graph metrics

• METRICFORENSICS intends to be scalable, effective and flexible.

• Highlights four interesting patterns in real-world volatile graphs:
  • Elbows
  • Broken correlations
  • Prolonged Spikes
  • Lightweight stars
METRICFORENSICS uses a suite of analysis techniques from:

- **Static graph mining**
  - But only it can drill down to various levels of detail

- **Dynamic graph mining**
  - Most works focus on evolution of communities and tensors
  - But they are limited to a single and fixed-time granularity as opposed to METRICFORENSICS

- **Anomaly detection**
  - Works are present like MDL, spectral analysis and OddBall.
  - All of these can be incorporated within METRICFORENSICS.

- **Time-series mining**
  - Techniques like BGP Lens can be brought under METRICFORENSICS’ fold.
The Drill-down

• METRICFORENSICS takes as input a volatile graph and determines its characteristic metrics.

• For a volatile graph, it is both difficult and time-consuming to perform various metric analysis continuously.

• So METRICFORENSICS rather focuses on doing sophisticated analysis only when interesting events occur.

• As the graph changes with time, it gives the ability to detect interesting events at various levels.
Continuing the Drill-down ... 

• Until an interesting event happens, METRICFORENSICS can collect graph global metrics like number of active node, number of active edges.

• The drill-down is performed on the ‘offending’ graph at which the interesting event is detected at more detailed topological and temporal resolutions.

• E.g., METRICFORENSICS can flag the sub-graph(s) that lead to the interesting turn of events and just focus on analyzing them.

• It can do community-level analyses like fraction of vertices in the community or tracking the changes in the structure of the community
Suppose it is a node in the community that interests us, METRICFORENSICS can track the change of its communities.

The drill-down can continue on to the complete local-level, i.e., just the node without taking into consideration its communities.

The local metrics that METRICFORENSICS can compute include:

- Centrality metrics
- Impact metrics
- OddBall metrics
  - OddBall metrics are used to find anomalies in weighted graphs.
  - OddBall analyzes the patterns of egonets to do this.
  - It is an unsupervised method that gives an ‘outlierness’ score to each and every node.
METRICFORENSICS Data Model

• METRICFORENSICS takes advantage of the fact that, for volatile graphs during a given time window, only a small percentage of the nodes and edges are active.

• **Snapshot graphs:**
  • For a time-step $t$, a snapshot graph $G_t$ is defined by its vertices $v_t$ and edges $e_t$ that are active at that moment.
  • Each time the graph structure changes, a new snapshot is generated.

• It is not feasible or worthwhile to consider just snapshot graphs.
Summary graphs

- A summary graph is a summary of the sequences of snapshot graphs during a time period T.

- Different ways for summarizing snapshot graphs:
  - Binary – An edge between nodes in summary graph exists between a pair of nodes if there ever existed any edge between them in one of the snapshot graphs in T.
  - Sum – Same as the above except for the fact that the weight of the edge in the summary graph is the sum of all existing edges between the pair in T.
  - Max – Similar to the above except that the weight is the maximum weight of all existing edges between a pair of nodes in T.

- Summary graphs can be generated either after a pre-defined number of time-steps or snapshot graphs.
More drilling

• METRICFORENSICS can also drill-down temporally.

• It can vary either the time intervals in which the
  • Summary graphs are analyzed: For streaming data, it maintains a circular buffer for recent snapshot graphs to drill-down on them when an interesting event occurs.
  • Summary graphs are generated: At a finer resolution, summary graphs are generated more frequently than coarser levels.

• Automation levels also vary in METRICFORENSICS.
More Analysis

- METRICFORENSICS also comprises of a suite of analysis techniques like:
  - Single Metric Analyses like:
    - Autoregressive Moving Average Model (ARMA).
    - Fourier Analysis
    - BGP-Lens
  - Coupled Metric Analysis
    - Correlation analysis between two metrics.
    - Summary graphs for which correlations between metrics are broken are deemed as interesting events.
    - Clustering like k-means can be performed for two time-series
  - Non-metric analysis
    - METRICFORENSICS has an innovative three-dimensional visualization
    - Inspection of attribute data
Experiments

• METRICFORENSICS was implemented in Java on an Intel Core 2 Duo machine.
• It attempted to answer the following questions:
  • *Can* METRICFORENSICS detect interesting events including anomalies?
  • *Do the discovered interesting events tell us something about volatile graphs themselves?*
  • *Is* METRICFORENSICS scalable and amenable to real-time execution?

• Volatile Graph Datasets:
  • An enterprise IP trace - **ENTP**
  • MIT Reality Mining proximity sensor data - **RMBT**
  • IP traffic from a specific enterprise port - **LBNL**
  • The largest graph has around 3 million nodes and more than thirty million edges.
Eigen Analysis

• Considers the relationship between the two largest eigenvalues $\lambda_1$ and $\lambda_2$ of the summary graph.

• Three types of edge-weighing strategies:
  • Maximum connections
  • Number of connections
  • Sum of bytes

• It has been observed that regardless of the weighing strategy, that there are “elbow regions” where the $\lambda_1$ is stable while $\lambda_2$ is changing and vice-verse.

• The principal eigenvalues of a weighted graph generally correspond with edge with the largest weight, the vertex with the largest degree or the component with the largest total weight.
Visualizing the “elbows”

\( \lambda_2 \) versus \( \lambda_1 \) under various edge-weighing strategies in the ENTP summary graphs (generated every 30 seconds). 

- \( x \)-axis is \( \lambda_1 \) in log-scale; \( y \)-axis is \( \lambda_2 \) in log-scale. The color of a dot is the time that it was observed (in minutes): blue is earlier, red is later. Regardless of the summary graphs’ edge-weighing strategy, there are interesting regions with elbow patterns where \( \lambda_1 \) is stable and \( \lambda_2 \) is changing, or vice versa.
LBNL elbow

$\lambda_2$ versus $\lambda_1$
Fractal Dimension Analysis

• Recall that the fractal dimensions indicate the ‘burstiness’ of a point collection.

• The fractal dimension $D$ of a point is zero while the fractal dimension of points uniformly distributed in a time-series is close to one.

• In the experiment, the fractal dimensions of different graph metrics on the summary graph were computed:
  • Window of 180 minutes on ENTP
  • Window of five minutes on LBNL
  • Window of ten days on RMBT

• Generally, $D$ here was observed to be stable and to be around 0.9 for ENTP/LBNL and 0.8 for RMBT.
The “prolonged spike”

# Additions
“Broken correlations”
“Light-weight stars”
Conclusion

- METRICFORENSICS is an efficient, scalable and flexible framework for analyzing volatile graphs with a good visual interface.
- It can observe interesting phenomena in volatile graphs like elbows, broken correlations, prolonged spikes and light-weight stars.
- It provides an innovative “drill-down” approach for analyzing volatile graphs at various temporal and topological levels.
- Extensively evaluated on real-world volatile graphs.
Higher-order Web Link Analysis Using Multilinear Algebra

KOLDA, BADER AND KENNY. ICDM 2005
Background - I

- Linear algebra techniques like PageRank (by Page & Brin) and HITS (by Kleinberg) are effective in web search.
- For a given query, HITS compiles a focused subgraph that has information more pertinent to the terms in the query.
- The main eigenvector of the adjacency matrix of this subgraph define the query’s best authorities and hubs.
- HITS suffers from the problem of “topic drift”.
Background - II

- HITS basically calculates an authority score \( (a) \) and a hub score \( (h) \) for every web page iteratively.

\[
h_i^{(t+1)} = \sum_{i \rightarrow j} a_j^{(t)} \quad \text{for} \quad i = 1, \ldots, n, \quad \text{and}
\]

\[
a_j^{(t+1)} = \sum_{i \rightarrow j} h_i^{(t+1)} \quad \text{for} \quad j = 1, \ldots, n,
\]

- Adjacency matrix representation of the graph:

\[
A_{ij} = \begin{cases} 
1 & \text{if } i \rightarrow j, \\
0 & \text{otherwise},
\end{cases}
\]

- SVD decomposition of \( A \):

\[
A \approx \sum_{i=1}^{p} \sigma^{(i)} u^{(i)} \circ v^{(i)}.
\]
Enter TOPHITS

• This paper introduces a new method called Topical Hypertext Induced Topic Selection (TOPHITS)

• TOPHITS does an analysis where the anchor text is combined with the hyperlink structure of the web to better facilitate search.

• The edges in the hyperlink graph are labelled with the anchor text.

• The limitation with Kleinber’s method is the adjacency matrix is not rich enough to hold the semantic anchor text information.

• Instead, a three-dimensional tensor is used to represent this ‘semantic’ graph.
Example Web

Endangered Species
Animals today are being threatened by a variety of environmental pressures. For example, the jaguar is losing prime habitat in the world. Zoos are trying to raise awareness of their plight.

Jaguar FAQ
Jaguars are an endangered species that live in the tropical rain forests of Central and South America. They live about 11 years in the wild and up to 22 years at a zoo.

Rain Forest Zoo
We have a new exhibit opening next month highlighting the endangered species of the Americas, including the jaguar.

Online Atlas
View maps of animal habitats from around the world, including those of endangered animals in North, South, and Central America.

Website 1
endangered

Website 2
jaguar

Website 3
zoo

Website 4
America

jaguars
species
endangered
zoo
America
The TOPHITS tensor

• The adjacency matrix A in HITS is just a tensor with two dimensions whereas the tensor in TOPHITS has three dimensions.

• The first two dimensions of the tensor are the indexes of the web pages/sites.

• The third dimension of the tensor are the topics themselves.

• The tensor is rich enough to capture the adjacency information for the Web for each and every topic separately.

• The topic information present in the tensor is what helps it avoid “topic drift” by incorporating ideas from Latent Semantic Analysis (LSA).
High-order representation example

A three-way tensor that models the semantic graph
PARAFAC

• Similar to SVD decomposition in HITS, TOPHITS uses Parallel Factors decomposition (PARAFAC) on the 3-D tensor. This gives the most significant factors of the tensor.

• More specifically, this produces triplets of vectors with authority and hub scores for web pages along with their topic scores.

• The adjacency tensor

\[ A_{ijk} = \begin{cases} 
1 & \text{if } i \rightarrow j \text{ with anchor text } k, \\
0 & \text{otherwise.} 
\end{cases} \]

• Decomposing this using PARAFAC gives,

\[ A \approx \sum_{i=1}^{p} \sigma^{(i)} u^{(i)} \circ v^{(i)} \circ w^{(i)} \]
HITS vs TOPHITS

In HITS model, the SVD provides a 2-way decomposition that yields authority and hub scores.

In TOPHITS, the PARAFAC model provides a 3-way decomposition that yields authority, hub, and topic scores.
Tophits iteration

\[
\begin{align*}
  h_i^{(t+1)} &= \sum_{i \rightarrow j} a_j^{(t)} t_k^{(t)} && \text{for } i = 1, \ldots, n, \\
  a_j^{(t+1)} &= \sum_{i \rightarrow j} h_i^{(t+1)} t_k^{(t)} && \text{for } j = 1, \ldots, n, \\
  t_k^{(t+1)} &= \sum_{i \rightarrow j} a_j^{(t+1)} h_i^{(t+1)} && \text{for } k = 1, \ldots, m.
\end{align*}
\]
Alternate expression

\[ h^{(t+1)} = A \times_2 a^{(t)} \times_3 t^{(t)}, \]

\[ a^{(t+1)} = A \times_1 h^{(t+1)} \times_3 t^{(t)}, \]

\[ t^{(t+1)} = A \times_1 h^{(t+1)} \times_2 a^{(t+1)} \]
TOPYHTS Algorithm

**In:** \( A \) of size \( n \times n \times m \).

**Out:** Rank-\( p \) approximation of \( A \), returned as \( p \) triplets \( \{ u^{(i)}, v^{(i)}, w^{(i)} \} \) plus weights \( \sigma^{(i)} \) for \( i = 1, \ldots, p \).

For \( k = 1, 2, \ldots, p \), do:

Initialize \( x, y, z \) to be vectors of all ones of length \( n, n, m \), resp.

Repeat:

\[
x = A \bar{x}_2 y \bar{x}_3 z - \sum_{i=1}^{k-1} \sigma^{(i)} u^{(i)} (y^T v^{(i)}) (z^T w^{(i)})
\]

\[
y = A \bar{x}_1 x \bar{x}_3 z - \sum_{i=1}^{k-1} \sigma^{(i)} v^{(i)} (x^T u^{(i)}) (z^T w^{(i)})
\]

\[
z = A \bar{x}_1 x \bar{x}_2 y - \sum_{i=1}^{k-1} \sigma^{(i)} w^{(i)} (x^T u^{(i)}) (y^T v^{(i)})
\]

\[
\lambda = \|x\|\|y\|\|z\|, \text{ and normalize } x, y, z
\]

Until the change in \( \lambda \) is small.

Set \( u^{(k)} = x, v^{(k)} = y, w^{(k)} = z, \sigma^{(k)} = \lambda \)

End do.
Evaluation – I

- The web-crawler was started from http://www-neos.mcs.anl.gov/neos
- Crawled around 4700 pages and 5600 cross-linked hosts.
- Only host-to-host links are considered and self-links are eliminated.
- Some cleansing is done to better extract information from pages
- Hyperlinks without anchor data are assigned the term “no-anchor-text”.
Evaluation – II

• The tensor is determined by counting the number of hyperlinks from host $i$ to host $j$ with term $k$

\[
A_{ijk} = \begin{cases} 
1 + \log(C_{ijk}) & \text{if } C_{ijk} \neq 0, \\
0 & \text{otherwise.}
\end{cases}
\]

• The standard adjacency matrix is also computed as per HITS and SVD is applied to it to get the hubs and authorities scores for it.

• The greedy PARAFAC algorithm was computed for the first twenty factors of the tensor.

• Cost of each iteration is $O(N)$, where $N$ is the number of nonzeros in the tensor.
Evaluation – III

• For each factor, TOPHITS gives a ranked list of hosts associated with a ranked list of terms.

• It includes terms that identify the topic of each authority sets.

• E.g., when searching for authorities on retirement, just search for the key term “retirement” to find the corresponding principal factor.

• Thus, TOPHITS is useful because it can automatically discover the topics along with the authorities/hubs for them.

• Thus it can work for multi-topic data sets.
### Results - I

<table>
<thead>
<tr>
<th>Topics</th>
<th>Term</th>
<th>Score</th>
<th>1st Principal Factor</th>
<th>Authority</th>
<th>Score</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.23</td>
<td>java</td>
<td>0.96</td>
<td></td>
<td>java.sun.com</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>0.18</td>
<td>sun</td>
<td>0.38</td>
<td></td>
<td>developers.sun.com</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>0.16</td>
<td>platform</td>
<td>0.46</td>
<td></td>
<td>doce.sun.com</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>0.16</td>
<td>solars</td>
<td>0.14</td>
<td></td>
<td>see.sun.com</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>0.16</td>
<td>developer</td>
<td>0.43</td>
<td></td>
<td><a href="http://www.sun.com">www.sun.com</a></td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>edition</td>
<td>0.09</td>
<td></td>
<td><a href="http://www.npa.gov">www.npa.gov</a></td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>download</td>
<td>0.07</td>
<td></td>
<td>developer.sun.com</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>0.14</td>
<td>info</td>
<td>0.06</td>
<td></td>
<td>sunsolve.sun.com</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>0.12</td>
<td>software</td>
<td>0.03</td>
<td></td>
<td>access1.sun.com</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>0.12</td>
<td>lehigh</td>
<td>0.06</td>
<td></td>
<td>lehigh.edu</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>lbib</td>
<td>0.06</td>
<td></td>
<td>www2.lehigh.edu</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>0.16</td>
<td>libraries</td>
<td>0.03</td>
<td></td>
<td>lehighalumni.com</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>lehigh</td>
<td>0.02</td>
<td></td>
<td>lehigh.edu</td>
<td>0.15</td>
<td></td>
</tr>
</tbody>
</table>

### 2nd Principal Factor

<table>
<thead>
<tr>
<th>Topics</th>
<th>Term</th>
<th>Score</th>
<th>1st Principal Factor</th>
<th>Authority</th>
<th>Score</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>lehigh</td>
<td>0.02</td>
<td></td>
<td><a href="http://www.ibm.com">www.ibm.com</a></td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>0.12</td>
<td>ibm</td>
<td>0.18</td>
<td></td>
<td><a href="http://www.alphaworks.ibm.com">www.alphaworks.ibm.com</a></td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>0.12</td>
<td>services</td>
<td>0.07</td>
<td></td>
<td><a href="http://www.ibm.com">www.ibm.com</a></td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>0.12</td>
<td>websphere</td>
<td>0.05</td>
<td></td>
<td><a href="http://www.developer.ibm.com">www.developer.ibm.com</a></td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>0.11</td>
<td>web</td>
<td>0.02</td>
<td></td>
<td><a href="http://www.redhouse.ibm.com">www.redhouse.ibm.com</a></td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>0.11</td>
<td>linux</td>
<td>0.01</td>
<td></td>
<td>research.ibm.com</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>0.11</td>
<td>resources</td>
<td>0.02</td>
<td></td>
<td><a href="http://www.redhouse.ibm.com">www.redhouse.ibm.com</a></td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>technologies</td>
<td>0.01</td>
<td></td>
<td><a href="http://www.redhouse.ibm.com">www.redhouse.ibm.com</a></td>
<td>0.10</td>
<td></td>
</tr>
</tbody>
</table>

### 3rd Principal Factor

<table>
<thead>
<tr>
<th>Topics</th>
<th>Term</th>
<th>Score</th>
<th>1st Principal Factor</th>
<th>Authority</th>
<th>Score</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.26</td>
<td>president</td>
<td>0.87</td>
<td></td>
<td><a href="http://www.whitehouse.gov">www.whitehouse.gov</a></td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>newshour-text</td>
<td>0.18</td>
<td></td>
<td><a href="http://www.whitehouse.gov">www.whitehouse.gov</a></td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>bush</td>
<td>0.16</td>
<td></td>
<td>travel.state.gov</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>0.17</td>
<td>welcome</td>
<td>0.10</td>
<td></td>
<td><a href="http://www.qsa.gov">www.qsa.gov</a></td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>0.16</td>
<td>white</td>
<td>0.08</td>
<td></td>
<td><a href="http://www.qsa.gov">www.qsa.gov</a></td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>0.16</td>
<td>us</td>
<td>0.08</td>
<td></td>
<td><a href="http://www.qsa.gov">www.qsa.gov</a></td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>0.13</td>
<td>budget</td>
<td>0.08</td>
<td></td>
<td><a href="http://www.whitehouse.gov">www.whitehouse.gov</a></td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>0.11</td>
<td>president</td>
<td>0.08</td>
<td></td>
<td><a href="http://www.whitehouse.gov">www.whitehouse.gov</a></td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>0.11</td>
<td>office</td>
<td>0.08</td>
<td></td>
<td><a href="http://www.whitehouse.gov">www.whitehouse.gov</a></td>
<td>0.11</td>
<td></td>
</tr>
</tbody>
</table>
Results - II

<table>
<thead>
<tr>
<th>Authority</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.ics.com">www.ics.com</a></td>
<td>0.97</td>
</tr>
<tr>
<td><a href="http://www.ics.com">www.ics.com</a></td>
<td>0.94</td>
</tr>
<tr>
<td><a href="http://www.ics.com">www.ics.com</a></td>
<td>0.92</td>
</tr>
<tr>
<td>1st Singular Vector</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.ics.com">www.ics.com</a></td>
<td>0.90</td>
</tr>
<tr>
<td><a href="http://www.ics.com">www.ics.com</a></td>
<td>0.88</td>
</tr>
<tr>
<td><a href="http://www.ics.com">www.ics.com</a></td>
<td>0.86</td>
</tr>
<tr>
<td>2nd Singular Vector</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.ics.com">www.ics.com</a></td>
<td>0.84</td>
</tr>
<tr>
<td><a href="http://www.ics.com">www.ics.com</a></td>
<td>0.82</td>
</tr>
<tr>
<td><a href="http://www.ics.com">www.ics.com</a></td>
<td>0.80</td>
</tr>
<tr>
<td>3rd Singular Vector</td>
<td></td>
</tr>
<tr>
<td>java.sun.com</td>
<td>0.80</td>
</tr>
<tr>
<td>java.sun.com</td>
<td>0.78</td>
</tr>
<tr>
<td>java.sun.com</td>
<td>0.76</td>
</tr>
<tr>
<td>4th Singular Vector</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.ics.com">www.ics.com</a></td>
<td>0.74</td>
</tr>
<tr>
<td><a href="http://www.ics.com">www.ics.com</a></td>
<td>0.72</td>
</tr>
<tr>
<td><a href="http://www.ics.com">www.ics.com</a></td>
<td>0.70</td>
</tr>
<tr>
<td>5th Singular Vector</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.ics.com">www.ics.com</a></td>
<td>0.68</td>
</tr>
<tr>
<td><a href="http://www.ics.com">www.ics.com</a></td>
<td>0.66</td>
</tr>
<tr>
<td><a href="http://www.ics.com">www.ics.com</a></td>
<td>0.64</td>
</tr>
</tbody>
</table>

HITS results
Results - III

Number of power method iterations per PARAFAC factor
Conclusion

• TOPHITS is a new approach for doing web search using higher-order presentations of the web and performing tensor decomposition on them.

• Determines the hubs and authorities for prominent topics in the web

• This method can deal with “topic drift” and work well for mult-topic datasets.
Questions