CIS 602-01: Scalable Data Analysis

Internet of Things

Dr. David Koop
Topological Data Analysis

• Properties
  - Coordinate Invariance
  - Deformation Invariance
  - Compressed Representation

• Fundamental Integration Tasks:
  - How does one infer high dimensional structure from low dimensional representations?
  - How does one assemble discrete points into global structure?

[Carlsson, Ghrist]
Topological Data Analysis Themes

• It is beneficial to replace a set of data points with a family of simplicial complexes, indexed by a proximity parameter. This converts the data set into global topological objects.

• It is beneficial to view these topological complexes through the lens of algebraic topology — specifically, via a novel theory of persistent homology adapted to parameterized families.

• It is beneficial to encode the persistent homology of a data set in the form of a parameterized version of a Betti number: a barcode

[Ghrist, 2008]
Points Cloud Set to a Complex

1. Which $\epsilon$? Converting a point cloud data set into a global complex (whether Rips, Čech, or other) requires a choice of parameter $\epsilon$. For $\epsilon$ sufficiently small, the complex is a discrete set; for $\epsilon$ sufficiently large, the complex is a single high-dimensional simplex. Is there an optimal choice for $\epsilon$ which best captures the topology of the data set? Consider the point cloud data set as a sequence of Rips complexes as illustrated in Figure 3. This point cloud is a sampling of points on a planar annulus. Can this be deduced? From the figure, it certainly appears as though an ideal choice of $\epsilon$, if it exists, is rare: by the time $\epsilon$ is increased so as to remove small holes from within the annulus, the large hole distinguishing the annulus from the disk is filled in.

2. Algebraic topology for data

Algebraic topology offers a mature set of tools for counting and collating holes and other topological features in spaces and maps between them. In the context of [Ghrist, 2008]...
Increasing the radius of the points

Figure 3. As $\epsilon$ increases, holes appear and disappear. Which holes are real and which are noise?

Despite being both computable and insightful, the homology of a complex associated to a point cloud at a particular $\epsilon$ is insufficient: it is a mistake to ask which value of $\epsilon$ is optimal. Nor does it suffice to know a simple 'count' of the number and types of holes appearing at each parameter value $\epsilon$. Betti numbers are not enough. One requires a means of declaring which holes are essential and which can be safely ignored. The standard topological constructs of homology and homotopy offer no such slack in their strident rigidity: a hole is a hole no matter how fragile or fine.

2.1. Persistence. Persistence, as introduced by Edelsbrunner, Letscher, and Zomorodian [12] and refined by Carlsson and Zomorodian [22], is a rigorous response to this problem. Given a parameterized family of spaces, those topological features which persist over a significant parameter range are to be considered as signal with short-lived features as noise. For a concrete example, assume that $R = (R_i)_{i=1}^N$ is as sequence of Rip complexes associated to a fixed point cloud for an increasing sequence of parameter values $(\epsilon_i)_{i=1}^N$. The natural inclusion maps $\iota : R_i \hookrightarrow R_{i+1}$ ensure 

$[\text{Ghrist, 2008}]$
Barcodes

For a finite persistence module $C$ with field $F$ coefficients,

$$H^\ast(C; F) \sim = \bigoplus_i x_t^i \cdot F[x] \oplus \left( \bigoplus_j x_r^j \cdot \left( F[x] / (x_s^j \cdot F[x]) \right) \right).$$

This classification theorem has a natural interpretation. The free portions of Equation (2.3) are in bijective correspondence with those homology generators which come into existence at parameter $t_i$ and which persist for all future parameter values. The torsional elements correspond to those homology generators which appear at parameter $r_j$ and disappear at parameter $r_j + s_j$. At the chain level, the Structure Theorem provides a birth-death pairing of generators of $C$ (excepting those that persist to infinity).

2.3. Barcodes. The parameter intervals arising from the basis for $H^\ast(C; F)$ in Equation (2.3) inspire a visual snapshot of $H_k(C; F)$ into a barcode. A barcode is a graphical representation of $H_k(C; F)$ as a collection of horizontal line segments in a plane whose horizontal axis corresponds to the parameter and whose vertical axis represents an (arbitrary) ordering of homology generators. Figure 4 gives an example of barcode representations of the homology of the sampling of points in an annulus from Figure 3 (illustrated in the case of a large number of parameter values $\epsilon_i$).

[Figure 4] An example of the barcodes for $H^\ast(R)$ in the example of Figure 3.

[Ghrist, 2008]
Data Polygamy

- Focus on urban data interactions
- Uncovering relationships between datasets aids in the understanding of cities
- Uncovering relationships → Uncovering important attributes
- Polygamy: many relationships
- Urban datasets can be very polygamous

[F. Chirigati et al., 2016]
Example: New York City

1. Would a reduction in traffic speed reduce the number of accidents?
2. Why is it so hard to find a taxi when it is raining?
3. Why is the number of taxi trips too low? Is this a data quality problem?

Hypothesis Testing vs. Hypothesis Generation

[Reference: F. Chirigati et al., 2016]
Defining Relationships between datasets

Challenges
1) How to define a relationship between data sets?

Hurricane Sandy
Hurricane Irene

[F. Chirigati et al., 2016]
Challenges & Solutions

• Defining relationships:
  - Finding interesting features
  - Accounting for both time and space
  - Solution: Topology-based relationships

• Large data complexity: Big urban data
  - 2.4 million possible relationships among NYC Open Data for a single spatio-temporal resolution
  - Data is at multiple resolutions
  - Solution: Hadoop

[F. Chirigati et al., 2016]
Topological Features: Merge Tree
Relationships

Positive

Negative

[F. Chirigati et al., 2016]
Framework and Scalability

Data Sets → Data Set Transformation → Feature Identification → Relationship Evaluation → Relationships

Performance Evaluation: Results
- Approach is robust to noise
- Approach is scalable

[F. Chirigati et al., 2016]
Projects

• Progress Report (Due Monday)
• Focus on scalability challenges or making something scalable
Internet of Things
What exactly is the "Internet of Things"?

Postscapes, Harbor Research
40 IoT Solutions

Vala Afshar
Internet of Things Technologies

- RFID
- (Wireless) sensor networks
- Addressing: IPv6
- Middleware: Object Abstraction, Service Management, Service Composition, Application

[Source: Botta et al., 2016]
Scalable IoT

• How to work with all of the data being generated by IoT?
  - Potential Answer: Cloud Technologies

• CloudIoT Drivers:
  - Communication
  - Storage
  - Computation
  - Scope

[Botta et al., 2016]
## Complimentary Aspects of the Cloud and IoT

<table>
<thead>
<tr>
<th></th>
<th>IoT</th>
<th>Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Displacement</td>
<td>Pervasive</td>
<td>Centralized</td>
</tr>
<tr>
<td>Reachability</td>
<td>Limited</td>
<td>Ubiquitous</td>
</tr>
<tr>
<td>Components</td>
<td>Real world things</td>
<td>Virtual resources</td>
</tr>
<tr>
<td>Computational</td>
<td>Limited</td>
<td>Virtually unlimited</td>
</tr>
<tr>
<td>capabilities</td>
<td></td>
<td>Virtually unlimited</td>
</tr>
<tr>
<td>Storage</td>
<td>Limited or none</td>
<td>Means for delivering services</td>
</tr>
<tr>
<td>Role of the Internet</td>
<td>Point of convergence</td>
<td>Means to manage</td>
</tr>
<tr>
<td>Big data</td>
<td>Source</td>
<td></td>
</tr>
</tbody>
</table>

[Source: Botta et al., 2016]
Intel's IoT Platform

The Intel® IoT Platform includes an end-to-end reference architecture and a portfolio of products from Intel and its ecosystem, that work with third-party solutions, to provide a foundation for seamlessly and securely connecting devices, delivering trusted data to the cloud, and delivering value through analytics.

**Intel® IoT Platform**
Secure, Scalable, Interoperable

- **SMART AND CONNECTED THINGS**
  Sense, filter, process, analyze, and actuate, while securing and managing machines and data.

- **DEVELOPER KITS, TOOLS & SDKs**
  Rapidly move to prototyping, piloting, and productizing with developer kits, tools, and SDKs.

- **UNLOCKING THE VALUE OF DATA**
  Process, store, and analyze data globally, perform complex analytics on large datasets, secure and manage millions of endpoints, and manage policies, metadata, and networks.

- **CONNECTING THE UNCONNECTED**
  Capture, filter, process, and store data, connect securely to legacy infrastructure, and perform analytics at the edge.

- **APIs AND THIRD-PARTY CLOUD CONNECTIONS**

- **DATA CENTER & STORAGE**

- **END-TO-END SECURITY**
  Secure hardware, software, and data, as well as device and policy management. Detect threats and safeguard scalable compute.

**VISUALIZE DATA AND MONETIZE INSIGHT**
Provide actionable information and create new services, while automating operations.
4. Applications

CloudIoT gave birth to a new set of smart services and applications, that can strongly impact everyday life (Fig. 5). Many of the applications described in the following (may) benefit from Machine-to-Machine communications (M2M) when the things need to exchange information among themselves and not only send them towards the cloud [42]. This represents one of the open issues in this field, as discussed in Section 7. In this section we describe the wide set of applications that are made possible or significantly improved thanks to the CloudIoT paradigm. For each application we point out the challenges – see Fig. 6 – which we discuss in detail in Section 5.

Healthcare.

The adoption of the CloudIoT paradigm in the healthcare field can bring several opportunities to medical IT, and experts believe that it can significantly improve healthcare services and contribute to its continuous and systematic innovation [43]. Indeed, CloudIoT employed in this scenario is able to simplify healthcare processes and allows to enhance the quality of the medical services by enabling the cooperation among the different entities involved. Ambient assisted living (AAL), in particular, aims at easing the daily lives of people with disabilities and chronic medical conditions. Through the application of CloudIoT in this field it is possible to supply many innovative services, such as: collecting patients’ vital data via a network of sensors connected to medical devices, [Botta et al., 2016]
Challenges for CloudIoT

<table>
<thead>
<tr>
<th>Applications</th>
<th>Privacy</th>
<th>Legal and social aspects</th>
<th>Large scale</th>
<th>Security</th>
<th>Reliability</th>
<th>Performance</th>
<th>Heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smart Home and Smart Metering</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Video Surveillance</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Healthcare</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Smart Cities and Communities</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Smart Energy and Smart Grid</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Automotive and Smart Mobility</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Smart Logistics</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Environmental Monitoring</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

- **Security and Privacy:**
  - Cryptography may not be possible at all layers due to power
  - Attack Examples: SQL Injection, Session Hijacking

- **Heterogeneity:** lots of different devices, OSes, platforms, services

- **Performance:** Network performance, real-time applications

- **Reliability:** e.g. smart cars

- **Legal and social aspects**

[Botta et al., 2016]
Challenges for CloudIoT

• Scalability: interaction of many devices across wide area
  - Storage
  - Computation
  - Monitoring

• Big Data:
  - Summarization
  - Databases
  - Data Integrity

[Botta et al., 2016]
Platforms, services, and projects

<table>
<thead>
<tr>
<th>Platform</th>
<th>Proprietary things</th>
<th>Open things</th>
<th>Private cloud</th>
<th>Public cloud</th>
<th>Free</th>
<th>Open source</th>
<th>Application API</th>
<th>Last update</th>
<th>Ready to use</th>
</tr>
</thead>
<tbody>
<tr>
<td>IoTCloud</td>
<td>✓</td>
<td>✓</td>
<td>n/a</td>
<td>n/a</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Oct. 14</td>
<td>✓</td>
</tr>
<tr>
<td>OpenIoT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Oct. 14</td>
<td>✓</td>
</tr>
<tr>
<td>IoT Toolkit</td>
<td>✓</td>
<td>✓</td>
<td>n/a</td>
<td>n/a</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Dec. 13</td>
<td>✓</td>
</tr>
<tr>
<td>NimBits</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>partly</td>
<td>✓</td>
<td>✓</td>
<td>n/a</td>
<td>Nov. 14</td>
<td>✓</td>
</tr>
<tr>
<td>openPicus</td>
<td>✓</td>
<td>✓</td>
<td>n/a</td>
<td>n/a</td>
<td>✓</td>
<td>✓</td>
<td>n/a</td>
<td>n/a</td>
<td>✓</td>
</tr>
<tr>
<td>Xively</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>n/a</td>
<td>✓</td>
</tr>
<tr>
<td>Open.Sen.se</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>n/a</td>
<td>✓</td>
</tr>
<tr>
<td>ThingSpeak</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>n/a</td>
<td>✓</td>
</tr>
<tr>
<td>CloudPlugs</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>n/a</td>
<td>✓</td>
</tr>
<tr>
<td>Carriots</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>n/a</td>
<td>✓</td>
</tr>
<tr>
<td>NetLab</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>n/a</td>
<td>Oct. 14</td>
<td>✓</td>
</tr>
<tr>
<td>Intel IoT Analytics</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Nov. 14</td>
<td>✓</td>
</tr>
<tr>
<td>Synapse IoT Cloud</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Nov. 14</td>
<td>✓</td>
</tr>
<tr>
<td>ClouT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>n/a</td>
<td>n/a</td>
<td>✓</td>
<td>n/a</td>
<td>✓</td>
</tr>
</tbody>
</table>

[Source: Botta et al., 2016]
Google Cloud
IoT and Big Data

G. Schmutz