Data analytics for Cyber security

- Anomaly Detection for cyber security-

Vandana P. Janeja

Outline

- What are anomalies
- Motivating example: BGP hijacking
- Challenges in understanding anomalies
  - Curse of dimensionality
  - Interpretation
  - Treating Anomalies

What are anomalies?

An anomaly in a data set is an observation of a set of observations that are different and inconsistent with respect to the normal observations

An anomaly is also referred to as outlier, peculiarity, exception, discord among others terms. Anomaly and outlier are mostly used interchangeably

An anomaly is always identified with respect to a frame of reference. The frame of reference is the normal or baseline set of observations in the data

An anomaly stands apart from this frame of reference

Anomaly detection heavily relies on discovering the normal, the frame of reference or the baseline against which an anomaly is compared, to be termed as an anomaly

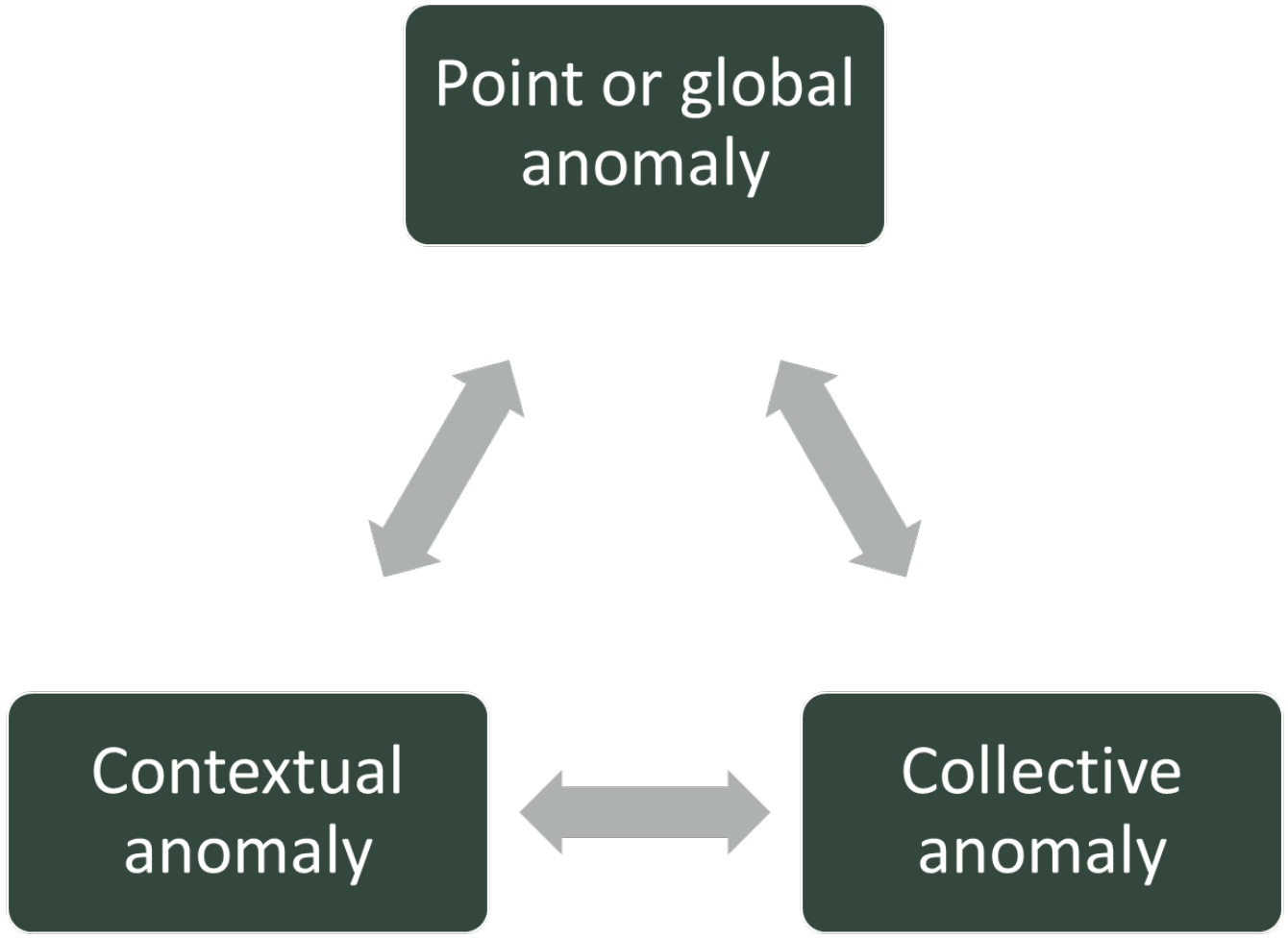
The further away the anomaly is from the frame of reference the more severe is the anomaly

Methods to detect Anomalies

- In the area of data analytics, anomaly detection has utilized techniques from
  - Supervised,
  - Semi supervised and
  - Unsupervised learning

Types of Anomalies

Anomalies have been characterized as single point, group of points and contextual anomalies, However, these different types of anomalies can be interrelated.



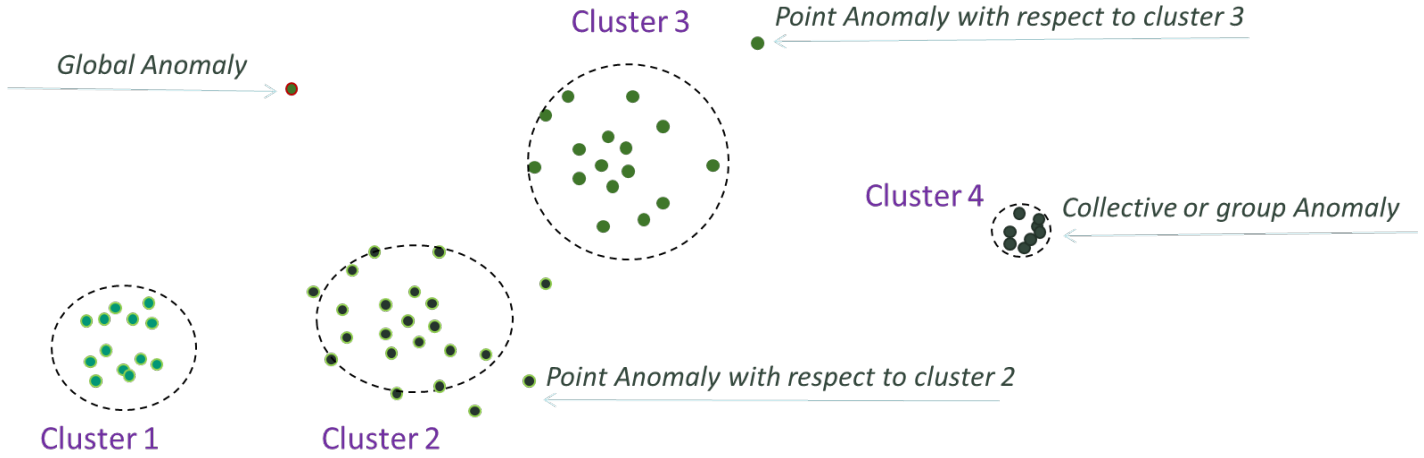
Types of Anomalies

- Point anomaly: is a single data point that is anomalous with respect to a group of neighboring points.
  - There can be different points off references such that an anomaly can be that respect to all of you in a budding points on a caster or in some cases the entire dataset.

- Collective or group Anomaly: A cluster is an anomaly with respect to all the clusters due to the relative size of the cluster and its distance from other clusters. Individual points inside the cluster may not be anomalous on their own, however, when considered together as a group they are anomalous.

- Contextual anomaly: is anomalous in the perspective of a certain context.
  - If the network traffic at a certain time appears to be very high as compared to the neighboring data points it may appear to be anomalous. However, if there is a specific event that is leading to the high traffic such as high sales during holiday time and then this high traffic would not be anomalous.
  - The context here is that the high amount of traffic is generated due to a marketing event or a special sale and not because of a flooding attack or a denial of service attack.
  - Defining a context can be very important to create a frame of reference, and can be done before the anomaly detection, which is much more robust, to clean up the data of possible frivolous outliers or as a post processing step where the outliers are filtered through any context.

Types of Anomalies



- Example of clusters of network traffic data which comprise of features such as time to live, datagram length, and total size.
- Once clustering is done we see that certain points do not fall into any of the clusters. We can evaluate this clustering as containing potential anomalies.
  - There is a single point that is a global anomaly which is an anomalous with respect to all the three clusters 1, 2, and 3. There some local anomalies which are anomalous with respect to specific clusters such as cluster two and cluster three.
- Cluster 4 as collective anomaly: In the example in figure, the point of reference is the other clusters, such that the size of the clusters 1, 2, 3 is much bigger and the distance of the small cluster 4 to the other clusters is much bigger than the distance between the clusters 1, 2, 3.
- Context: cluster 4 is in fact a group of points where a large file was being uploaded by a specific network admin, which resulted in massive size transfer on the network generating the data points that are different than the rest of the traffic.
- Given the context of the origin of the data this data should not have been included in the analysis.
- A frame of reference is extremely important in discovering anomalies such that points or groups of points are not frivolously identified as anomalous.

Anomaly - Context

Context can be in the form of attributes within the data that can be used to create demarcations in the data so that they are separated into groups for further analysis.

Context can help define homogenous groups for anomaly detection in them.

Context discovery as a preprocessing step where the groups are identified using contextual attributes and then anomaly detection is performed in the contextually relevant groups of data points using the their behavioral attributes.

Context aware computing allows for accommodating context definition into the algorithmic process.

Data points with similar context should be grouped together and anomaly detection can be applied to each of the groups for a well-defined anomaly detection.

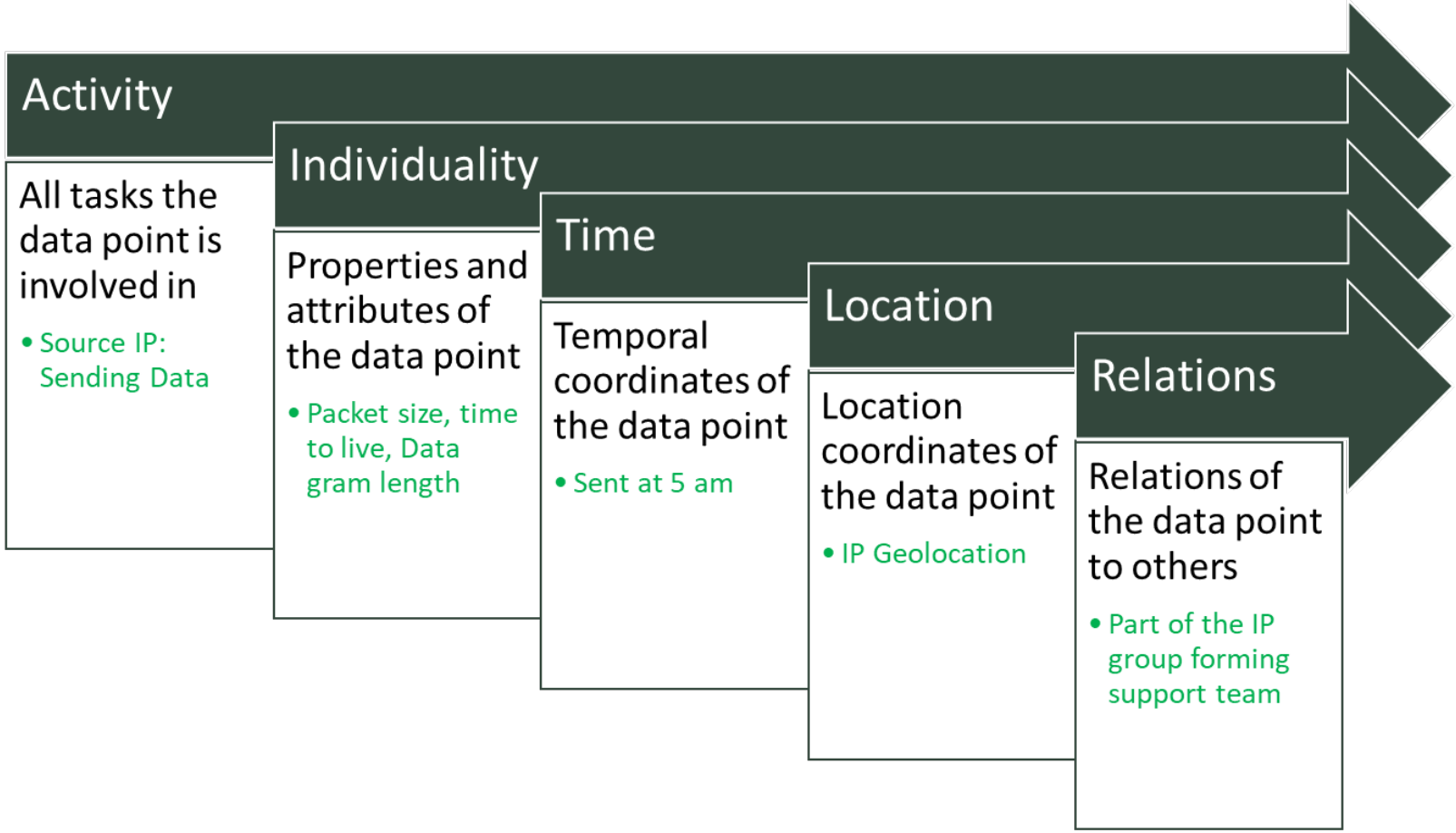
Point, group and contextual anomalies are related to each other since a context can apply to both point and groups of outliers.

In some cases individual point anomalies can collectively be part of a group anomaly, this may not always be the case since data points which are part of the group may or may not be individually anomalous.

Anomaly detection has several associated challenges, these will be clarified next with the help of a motivating example.

Anomaly-Context

- Context can be defined in the terms of activity, individuality, spatial or location context, temporal context and relations context
- For example the data point is a source and is sending the data packet at 5 am and is part of the IP address allocated to the support team.
- If all the support team IP's are sending similarly sized data and at similar time frame then this point will be considered normal.
- However if this particular data point is sending data at an unusual time and is originating form a geolocation outside of the expected locations then this could be a potential anomaly.
- Here the frame of reference is defined in terms of time, IP location and the team.



Motivating Example: BGP Hijacking

Network packets follow certain specific traffic routes on the internet using the announced and available routes maintained in the routing tables

Border Gateway Protocol (BGP) is primarily responsible for exchange of information between Autonomous Systems (AS) for successful transmission over the published routes

Several types of attacks lead to hijacking of hosts or servers and redirecting traffic to anomalous sites or dumping traffic at random sites

BGP hijacking is one such type of attack, examples include: redirection of Google and you tube requests in Turkey and, erroneous rerouting of YouTube traffic among others

Motivating Example: BGP Hijacking

- BGP anomalies prevent successful exchanges between the AS leading to loss of data or unauthorized rerouting of the data
- Anomalies may include route flapping, path announcements that delay routed packets, malicious redirecting of traffic for unauthorized surveillance purposes, and dropping of packets to unknown destinations
- BGP anomalies where path announcements are made may lead to thousands of anomalous updates
- BGP first initiates a complete routing table exchange and then subsequently exchanges updates
- If there is an anomaly then there will be a lot more updates than normal. This can potentially indicate the presence of an anomaly
- A more specific discovery is needed to identify where (at what location or specific AS) did the anomaly originate as this can provide pertinent information for interpretation of the anomaly whether it was a simple misconfiguration or a more malicious intent to disrupt or redirect the traffic

Motivating Example: BGP Hijacking

- BGP anomalies can also be classified similar to the general anomalies a point or collective anomaly
- A single update is considered an (point) anomaly if it has an invalid AS number, invalid IP prefix, or a prefix announced by an illegitimate AS
- A set of updates can be anomalous (collectively) if there are several BGP updates in a short amount of time containing longest and shortest paths or substantial changes in the BGP traffic
- Such anomaly detection can use data from BGP update messages or routes that packets use

BGP data repositories

- Several repositories are available for BGP datasets to be analyzed and for discovering anomalies
- Each of these datasets shows the complexity and massive nature of the data.

Data Description	Link
RouteViews	http://www.routeviews.org/
Reseaux IP Europeens	https://www.ripe.net/manage-ips-and-asns/db
BGPMon	https://bgpmon.net/
Internet route registries	http://irr.net/index.html
CIDR report (bogon prefix lists)	http://www.cidr-report.org/as2.0/
MaxMind, GeoIP location data	https://www.maxmind.com/en/open-source-data-and-api-for-ip-geolocation
IP2location	http://www.ip2location.com/
CAIDA traceroute platform	http://www.caida.org/projects/ark/

Challenges in understanding anomalies

Curse of dimensionality

Interpretation

Treating Anomalies

Curse of dimensionality

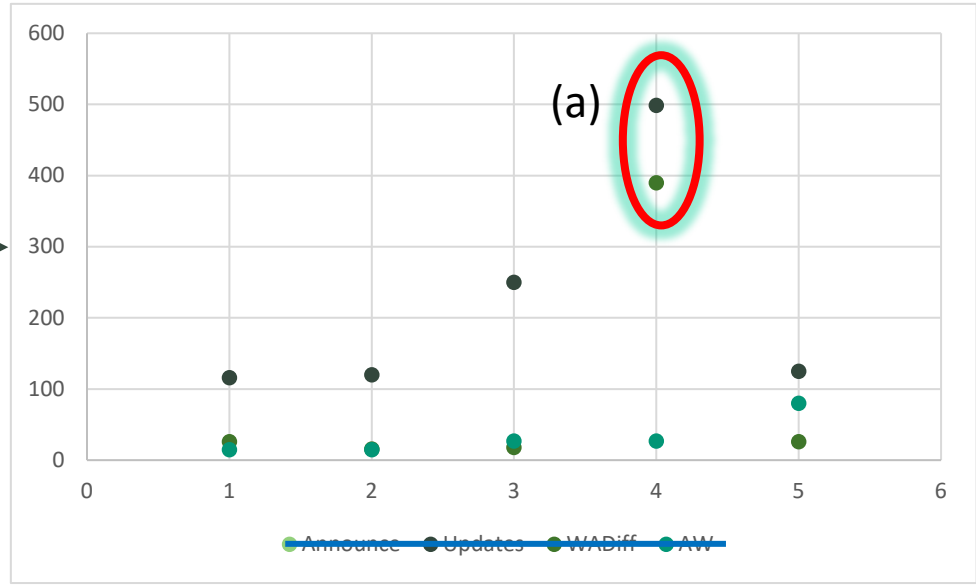
- As data size increases both in terms of the number of data points and the features describing them the patterns in the data get further embedded deeper into the data space, primarily because data is sparse in high dimensions.
- Patterns may exist in some sub spaces in the data and not in others
- Referred to as the curse of dimensionality
- Curse of dimensionality affects all the data analytics techniques and is even more relevant for anomaly detection as anomalies are already rare
- High dimensional data can be analyzed in various ways to identify anomalous routes
- The full dataset can be considered for the analysis or alternatively a smaller relevant subset of the data can be considered
- In a BGP dataset: It is possible that a route or a single update can be an outlier in one subset of attributes but not an outlier in another subset, For example: a route is not an outlier in the update attribute but is an outlier in the location attribute.
- It is essential to carefully determine whether all the dimensions are to be considered or a selected critical attribute set is considered
- Selection of the critical attribute set is also a data-mining task in itself such as through feature selection

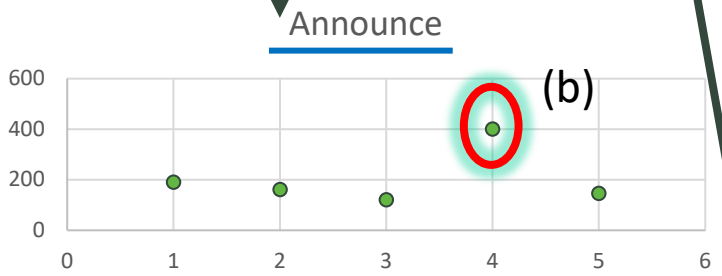
Example- Curse of Dimensionality

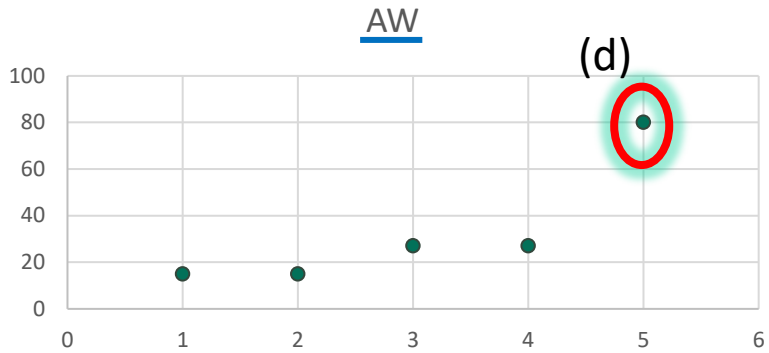
- The data is comprised of four attributes: number of BGP announcements, number of BGP updates, number of new paths announced and number of withdrawals after announcing the same path are used
- The two-dimensional plot of the data considering all the attributes together where (a) depicts clear anomalies which stand out from the other data points
- This anomaly corresponds to the observation 4 which shows high values for all attributes with respect to all the other observations and b, c, d, e show the plot for each attribute
- Observation 4 dominates plots in b, c, and d
- In (c) we see that observation 3 which has a high number of BGP updates is also deviant with respect to the rest of the data points
- In (d) which plots the attribute AW i.e, the number of withdrawals after announcing the paths, it is evident that observation 5 is highly deviant as compared to the rest of the data points
- This anomaly did not show up in (a) and was masked by the other highly deviant values
- Depicts the challenge of high dimensionality when the patterns are hidden in the subspaces

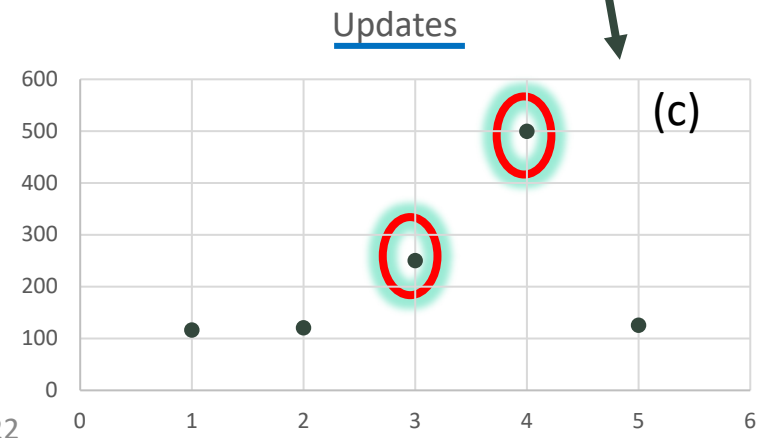
- This type of analysis based on a clearly defined context could be useful in selecting the suspicious routes for further investigation. Such analysis can help answer:
  - What routes or updates should a point be compared against?
  - What are the critical dimensions, which facilitate the discovery of anomalous routes?
  - Why is a certain route an outlier, namely identifying the dimensions causing the outlierness?
  - If all dimensions are being used for the discovery then to account for contribution from all dimensions and also identify the (outlier) causal dimension.

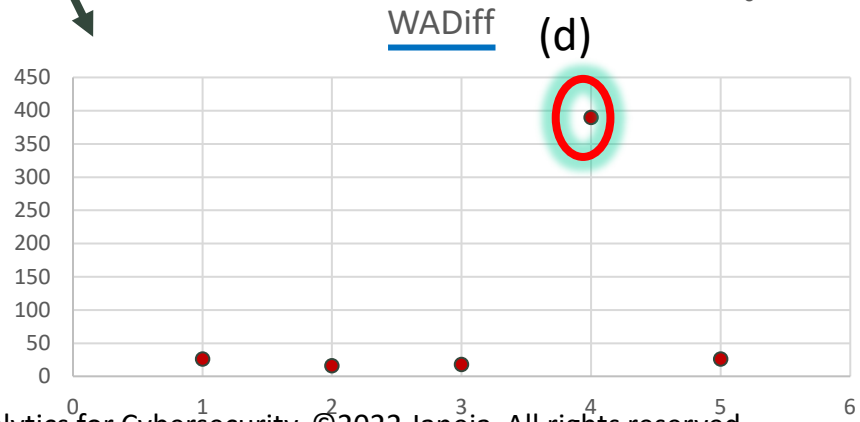
Observations	Announce	Updates	WADiff	AW
	(#of BGP Announcements)	(#BGP updates)	# of new paths announced after withdrawing an old path	# of withdrawals after announcing the same path
1	190	116	26	15
2	160	120	16	15
3	120	250	18	27
4	400	499	390	27
5	145	125	26	80







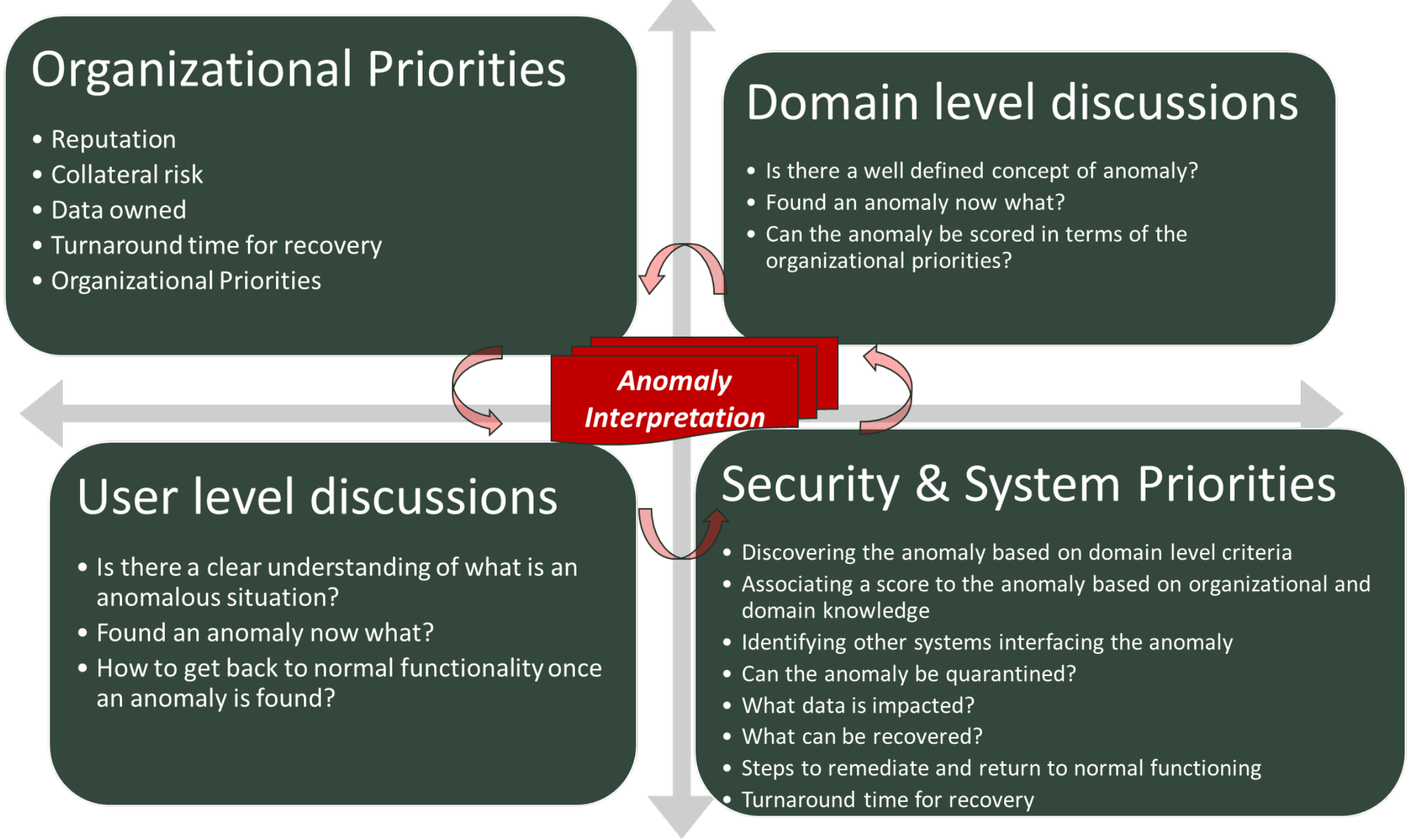




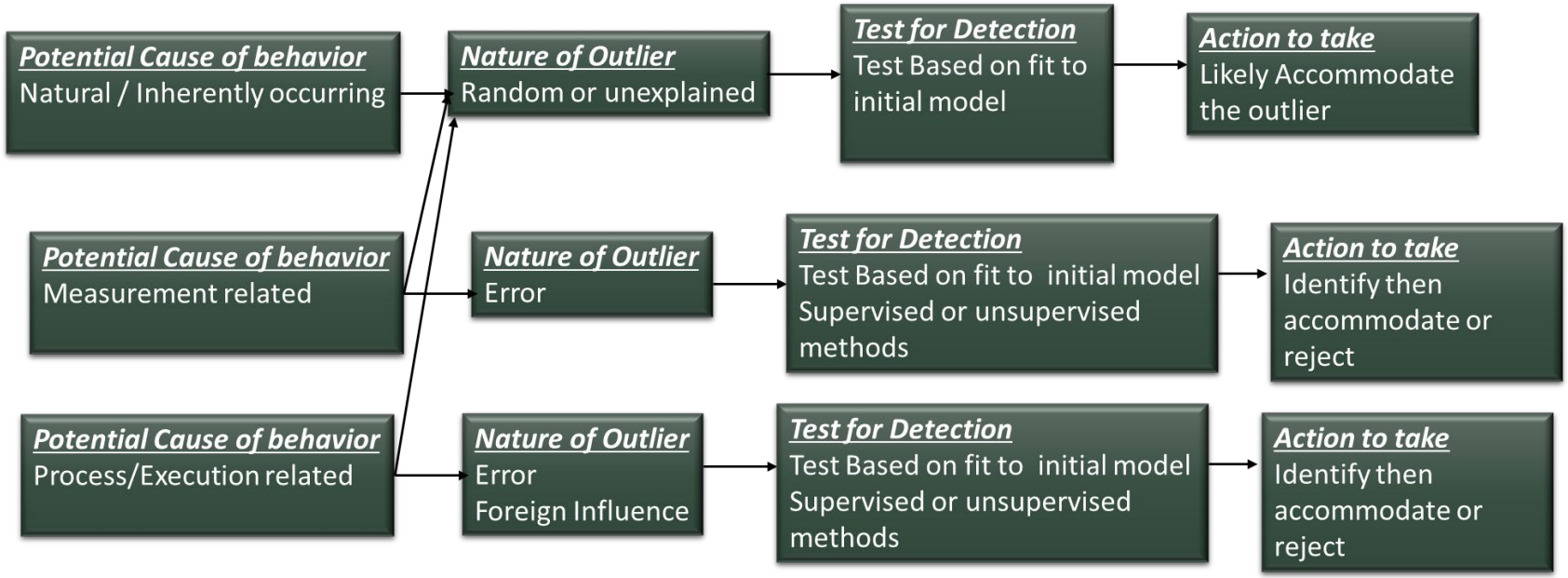
Example-
Curse of Dimensionality

Interpretation of Anomalies

- Anomalies can be a nuisance for an operator or a user who does not want too many alerts to slow them down
- Too many alerts can lead to Alert fatigue, desensitizing the end users to the numerous pings from their systems
- This can affect the overall performance of a system but more importantly this can desensitize users to click when they should not be clicking
- For a security officer in a complex organization each anomaly may be a critical alert to portend the arrival of a major attack or a major misconfiguration leaving a gaping hole in a system
- Priorities and discussions at various levels can impact the discovery and interpretation of an anomaly
- Interpretation of anomalies will vary from one type of user to another, one type of system to another and one domain to another based on the context defined and agreed upon by each
- Users in the same organization may also not agree on the definition of an anomaly, making it critical to have a well-crafted definition of an anomaly to
  - facilitate the discovery of an anomaly,
  - reduce the false positives and
  - associate importance to the level of the anomaly for faster recovery and remediation



Treating Anomalies



Treating Anomalies- Accommodation

- Accommodation tends to detect the outlier and accommodate the value in the data set so that the whole set of observations are still intact
- Accommodation will manipulate the observation so that even if an anomaly is in the dataset, the data does not become skewed because of this rogue observation
- Accommodation treats outliers as part of the data, although their removal fits the data into some distribution or normalizes the data, their presence could be equally important for the analysis
- The analysis can still be safely performed without any loss of data or observations, aiming at the robustness of the analysis in presence of outliers

Treating Anomalies- Rejection

- Rejection is testing outlier purely from the point of view of rejecting it from the data set or of identifying it as a feature of special interest and removing it before further analysis is done

- If the anomaly is cased due to a measurement error it is either rejected, corrected or the detection is repeated
- If the variation is inherent or due to some real measurement of an event then the anomaly detection is used but again in this case the outlier is incorporated in the revised model, identified for a separate study of origin and form
- If the anomaly is due to some random reason then it could be accommodated to see the overall analysis rather than eliminating the observation altogether.

Treating Anomalies-Cybersecurity

- In terms of cybersecurity, majority of the cases the anomaly must be eliminated.
  - If there is a Trojan discovered in a computer system which disrupts the regular functionality it has to be eliminated
  - If the BGP updates appear to be anomalous they need to be corrected for proper communication and movement of the traffic.
  - In some cases the anomaly needs to be studied in which case the anomaly may be detected and traced.
  - Example: “Jail” where an attacker can be traced through the system access and observations collected to learn about the role and activities of the attacker
  - Example: “honeypots” are set up to collect data and activities of attackers

References

- V. Barnett and R. Lewis. Outliers in statistical data. John Wiley and Sons, 1994
- Al-Musawi, B., Branch, P., & Armitage, G. (2016). BGP Anomaly Detection Techniques: A Survey. *IEEE Communications Surveys & Tutorials*.
- Pierluigi Paganini, 2014, Turkish Government is hijacking the IP for popular DNS providers
- http://securityaffairs.co/wordpress/23565/intelligence/turkish-government-hijacking-dns.html, Last accessed June 2017
- Greg Sandoval, 2008, YouTube blames Pakistan network for 2-hour outage https://www.cnet.com/news/youtube-blames-pakistan-network-for-2-hour-outage/ , Last accessed June 2017
- Wübbeling, M., Elsner, T., & Meier, M. (2014, June). Inter-AS routing anomalies: Improved detection and classification. In *Cyber Conflict (CyCon 2014), 2014 6th International Conference On* (pp. 223-238). IEEE.
- Al-Musawi, B., Branch, P., & Armitage, G. (2015, December). Detecting BGP instability using Recurrence Quantification Analysis (RQA). In *Computing and Communications Conference (IPCCC), 2015 IEEE 34th International Performance* (pp. 1-8). IEEE.
- Schlamp, J., Carle, G., & Biersack, E. W. (2013). A forensic case study on as hijacking: The attacker's perspective. *ACM SIGCOMM Computer Communication Review*, *43*(2), 5-12.
- Li, J., Dou, D., Wu, Z., Kim, S., & Agarwal, V. (2005). An Internet routing forensics framework for discovering rules of abnormal BGP events. *ACM SIGCOMM Computer Communication Review*, *35*(5), 55-66.
- Al-Rousan, N. M., & Trajković, L. (2012, June). Machine learning models for classification of BGP anomalies. In *High Performance Switching and Routing (HPSR), 2012 IEEE 13th* International Conference on (pp. 103-108). IEEE.
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15.
- Patcha, A., & Park, J. M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer networks, 51(12), 3448-3470.
- Estevez-Tapiador, J. M., Garcia-Teodoro, P., & Diaz-Verdejo, J. E. (2004). Anomaly detection methods in wired networks: a survey and taxonomy. Computer Communications, 27(16), 1569-1584.
- Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19-31.
- Ten, C. W., Hong, J., & Liu, C. C. (2011). Anomaly detection for cybersecurity of the substations. *IEEE Transactions on Smart Grid*, *2*(4), 865-873.
- Hong, J., Liu, C. C., & Govindarasu, M. (2014). Integrated anomaly detection for cyber security of the substations. *IEEE Transactions on Smart Grid*, *5*(4), 1643-1653.
- Feily, M., Shahrestani, A., & Ramadass, S. (2009, June). A survey of botnet and botnet detection. In *Emerging Security Information, Systems and Technologies, 2009. SECURWARE'09. Third International Conference on* (pp. 268-273). IEEE.
- Zimmermann, A., Lorenz, A., & Oppermann, R. (2007, August). An operational definition of context. In *International and Interdisciplinary Conference on Modeling and Using Context* (pp. 558-571). Springer Berlin Heidelberg.
- R. Bellman, Adaptive Control Processes: A Guided Tour, Princeton University Press, Princeton, New Jersey (1961).
- Doug Drinkwater, 2016, Does a data breach really affect your firm's reputation? http://www.csoonline.com/article/3019283/data-breach/does-a-data-breach-really-affect-your-firm-s-reputation.html, Last accessed June 2017
- Cheswick, B. (1992, January). An Evening with Berferd in which a cracker is Lured, Endured, and Studied. In *Proc. Winter USENIX Conference, San Francisco* (pp. 20-24).
- Spitzner, L. (2003). *Honeypots: tracking hackers* (Vol. 1). Reading: Addison-Wesley.