Chapter 1
Reliable Aggregation on Network Traffic for Web Based Knowledge Discovery

Shui Yu, Simon James, Yonghong Tian and Wanchun Dou

Abstract The web is a rich resource for information discovery, as a result web mining is a hot topic. However, a reliable mining result depends on the reliability of the data set. For every single second, the web generate huge amount of data, such as web page requests, file transportation. The data reflect human behavior in the cyber space and therefore valuable for our analysis in various disciplines, e.g. social science, network security. How to deposit the data is a challenge. An usual strategy is to save the abstract of the data, such as using aggregation functions to preserve the features of the original data with much smaller space. A key problem, however is that such information can be distorted by the presence of illegitimate traffic, e.g. botnet recruitment scanning, DDoS attack traffic, etc. An important consideration in web related knowledge discovery then is the robustness of the aggregation method, which in turn may be affected by the reliability of network traffic data. In this chapter, we first present the methods of aggregation functions, and then we employe information distances to filter out anomaly data as a preparation for web data mining.

Shui Yu
School of Information Technology, Deakin University, Victoria, Australia, e-mail: syu@deakin.edu.au.

Simon James
School of Information Technology, Deakin University, Victoria, Australia, e-mail: sjames@deakin.edu.au.

Yonghong Tian
School of Electronic Engineering and Computer Science, Peking University, Beijing, China, e-mail: yhtian@pku.edu.cn.

Wanchun Dou
Department of Computer Science and Technology, Nanjing University, Nanjing, China, e-mail: douwc@nju.edu.cn

1.1 Introduction

Web based data mining is a very practical and promising field for knowledge discovery, and plenty work has been done in this area [Srivastava et al., 2000] [Cooley, 2000] [Manavoglu et al., 2003]. Due to its size and complexity, the Internet is an invaluable resource when it comes to the study of networks, their underlying structure and features. More than just an information repository, understanding user interactions across the Internet can also provide insights into the behavior of social networks and ecological systems.

Faced with such massive amounts of information, it is helpful to aggregate network traffic data into a single value which represents a particular state of flow. In statistics, the most common way to aggregate information is to use the arithmetic mean, or in the case of skewed distributions or outliers, the median. There are many other aggregation functions, however which can provide a more reliable evaluation of the data. In particular, the ordered weighted averaging (OWA) function defined by Yager in [Yager, 1988] and its extensions (See [Yager & Beliakov, 2010] for a recent overview). The OWA generalizes both the median and arithmetic mean, and with an appropriate selection of parameters is able to aggregate an arbitrary number of "normal" arguments and discard outliers.

An important extension of the OWA function is the induced OWA, which orders its arguments using an auxiliary variable associated with the arguments rather than the arguments themselves. Distance-based aggregation or classification such as k-nearest neighbors, time-series smoothing and group decision making taking expertise into account can all be framed in terms of an induced OWA. Of interest to us is the ability of the induced OWA to reliably aggregate network traffic information, even in the presence of abnormal or attack data. To do this however, the induced OWA relies on an evaluation of similarity or distance between inputs, which in the case of Internet traffic will often take the form of distributions.

One of the challenges for reliable aggregation can then be framed in terms of evaluating the similarity between the current flow of traffic and legitimate user activity. Some of the common measures of similarity used information theory were recently compared toward this end in [Yu et al., 2009]. Once we are able to determine the reliability of information, we can then summarize it - either using a function like the OWA or another function if we remove the misleading data first.

In Section 1.2 we briefly introduce the motivation for the use of similarity- and distance-based aggregation functions, namely the threat of DDoS attacks and how this creates problems when it comes to the use of Internet traffic information in data analysis. We then give a brief overview of aggregation functions with a particular focus on the OWA and its extensions in Section 1.3. Some common measures of similarity drawn from information theory will be presented in Section 1.4, which is followed by comparisons of the information distances in Section 1.5. The last section summarizes the chapter.

1.2 The Reliability of Network Traffic Information

The enormous benefit provided to users as a result of the Internet's scale, and the ease with which information can be accessed and distributed comes at a price. Users and information providers can be vulnerable to viruses, security threats, and increasingly over the last decade - attacks designed to disrupt service.

Internet attacks with the aim of crashing a server or website so that legitimate users cannot access the information are known as Denial of Service (DoS) attacks. Peng et al. conducted a survey of such attacks and defense methods [Peng et al., 2007] [Thing et al., 2007], noting that a key challenge for defense was to be able to better discriminate between legitimate and malicious information requests. They note the exponential growth of attacks, with less than 20,000 reported in 2000 going up to almost 140,000 in 2003. It is further noted that the potential threat posed by DoS attacks may not just be in monetary terms but also to human life, as more and more emergency services come to rely on the Internet.

While some DoS attacks aim to crash systems by exploiting vulnerabilities in software code or protocols, the more prevalent form in recent years have been those which attempt to overload the resources of the server computer with fake or useless requests in huge volumes. These attacks can be more difficult to guard against since the vulnerability lies in physical bandwidth resources rather than an operating flaw, and the requests generated by the attacker may effectively mimic that of real users. A common strategy for delivering such attacks is for the attacker to take control of other computers, compromising their systems and distributing the attacks from up to 100s of 1000s of sources. These attacks are known as Distributed Denial of Service (DDoS) attacks [Wang et al., 2007] [Moore et al., 2006] [El-Atawy et al., 2009].

In the case of DDoS attacks, distinguishing between legitimate and false requests at the packet level is impossible. One approach is to to distinguish attacks based on the flow of traffic at any point in time. The ability to do this has obvious benefits for defense against such attacks, however we can also see the benefit to assessments of the reliability of network traffic information. Since traffic flow can tell us so much about the Internet and other networks, we want to ensure that we have an accurate picture of user behavior and the relationship between nodes, and that this picture is not distorted by the presence of illegitimate data packets.

In the following sections, we investigate the problem of evaluating the reliability of network traffic information. We look to identify legitimate data based on the similarity of its distribution with the distribution patterns present in the case of usual traffic.

1.3 Aggregation Functions

The need to aggregation information into a single numerical descriptor arises naturally in various fields. In statistics, the most commonly employed aggregation functions are the arithmetic mean and the median. Recent books concerning aggregation

functions include [Beliakov et al., 2007, Grabisch et al., 2009, Torra & Narukawa, 2007]. The following definitions will be useful.

Definition 1. A function $f : [0,1]^n \to [0,1]$ is called an aggregation function if it is monotone non-decreasing in each variable and satisfies $f(0,0,\ldots,0) = 0$, $f(1,1,\ldots,1) = 1$.

An aggregation function is referred to as *averaging* if the output is bounded by the maximum and minimum input. The arithmetic mean and median are examples of this type of aggregation. In their standard form, both these functions treat each input with equal importance. Weights, usually denoted w_i can be associated with each input, and where the weights are non-negative and satisfy $\sum_{i=1}^{n} w_i = 1$, the functions will remain averaging.

The weighted arithmetic mean is then the function

$$M_{\mathbf{w}}(\mathbf{x}) = w_1x_1 + w_2x_2 + \ldots + w_nx_n = \sum_{i=1}^{n} w_ix_i. \tag{1.1}$$

Rather than basing the importance on the source of the input, the ordered weighted averaging function assigns its weights based on their relative magnitude. It is given by

$$OWA_{\mathbf{w}}(x_1,\ldots,x_n) = \sum_{i=1}^{n} w_ix_{(i)}, \tag{1.2}$$

where the $(.)$ notation denotes the components of $\mathbf{x}$ being arranged in non-increasing order $x_{(1)} \geq x_{(2)} \geq \ldots \geq x_{(n)}$. The OWA is capable of expressing a number of order statistics such as the maximum function where $\mathbf{w} = (1,0,...,0)$ and the minimum for $\mathbf{w} = (0,...,0,1)$. It is also convenient for expressing the median $\mathbf{w}_k = 1$, for $n = 2k+1$ (n is odd) or $w_k = w_{k+1} = 0.5$ for $n = 2k$ (n is even) and $w_i = 0$ otherwise.

The OWA was used in [Yager & Beliakov, 2010] to replace the sum of least squares in regression problems. Reliable linear models were hence fit to the data which did not consider abnormal or extreme values in the determination of the weights. Similar goals could be sought in web-knowledge discovery, where we want to understand and interpret the data, however we do not want these findings to be distorted by the presence of attack traffic. Sometimes however, we will be looking to aggregate information which may not be extreme itself, but may be generated by illegitimate traffic. With the reordering step, the OWA's behavior differs on different parts of the domain. The induced OWA provides a more general framework for this reordering process. An inducing variable can be defined, over either numerical or ordinal spaces, which then dictates the order by which the arguments are arranged.

The induced OWA as stated by Yager and Filev in [Yager & Filev, 1999] is given by

$$IOWA_{\mathbf{w}}(\langle x_1,z_1\rangle,\ldots,\langle x_n,z_n\rangle) = \sum_{i=1}^{n} w_ix_{(i)}, \tag{1.3}$$

where the $(.)$ notation now denotes the inputs $\langle x_i,z_i\rangle$ reordered such that $z_{(1)} \geq z_{(2)} \geq \ldots \geq z_{(n)}$ and the convention that if q of the $z_{(i)}$ are tied,

i.e. $z_{(i)} = z_{(i+1)} = \ldots = z_{(i+q-1)}$,

$$x_{\eta(i)} = \frac{1}{q} \sum_{j=\eta(i)}^{\eta(i+q-1)} x_j,$$

Where x_i provides information to be aggregated, z_i provides some information about x_i, e.g. the importance, distance from the source, time displacement of the reading etc. The input pairs $\langle x_i, z_i \rangle$ may be two independent features of the same input, or can be related by some function.

Example 1. For the weighting vector $\mathbf{w} = (0.6, 0.3, 0.1)$, and the input $\langle \mathbf{x}, \mathbf{z} \rangle = (\langle 0.2, 3 \rangle, \langle 0.7, 2 \rangle, \langle 0.05, 8 \rangle)$, the aggregated value for the induced OWA is

$$IOWA_{\mathbf{w}}(\langle x, z \rangle) = 0.6(0.05) + 0.3(0.2) + 0.1(0.7) = 0.16 \,.$$

In the following section, we have in mind the use of similarity between legitimate and attack traffic as our auxiliary variable. We can then allocate less or no weight to the traffic information which differs greatly to the usual flow patterns.

1.4 Information Theoretical Notions of Distance

To evaluate the reliability of web traffic data, we need to determine whether or not the current flow is consistent with legitimate activity. In other words, we are interested in how similar the current traffic pattern is to the normal state. Evaluating this similarity can be problematic in the case of DDoS attacks, since the aim of the attack is to overwhelm the victim with the sheer number of data packets and hence their content need not differ at all to legitimate requests. One approach is to consider the characteristics of the traffic's distribution and use these to form the basis of similarity assessments. The dominating characteristic of a DDoS attack is clearly the abnormally high volume of traffic, however this feature alone makes attacks difficult to distinguish from flash crowds, which are generated by legitimate users [Yu et al., 2008]. A key problem in assessing the reliability of data concerning network behavior is then the ability to discern between authentic packet flows, which may vary and include periods of high congestion as well as base low-level volumes, and traffic rates due to malicious attacks or other noise which could affect methods of knowledge discovery.

Mathematical notions of similarity differ depending on the application. In classification and statistical analysis, similarity between multi-variate objects is often defined in terms of Pearson's correlation coefficient or the cosine of the angle between their vectors. Similarity is also often interpreted in terms of distance, which once again can be measured in a number of different ways, usually using some metric. A metric is a function of two objects $d(\mathbf{x}, \mathbf{y})$ (which may be single- or multi-variate), which is non-negative and satisfies the following conditions:

$$1)\ d(\mathbf{x},\mathbf{y}) = 0 \Longleftrightarrow \mathbf{x} = \mathbf{y};$$
$$2)\ d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x});$$
$$3)\ d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z}) \geq d(\mathbf{x},\mathbf{z}).$$

Condition 2 ensures that the metric is symmetric, i.e. the distance between two objects is not dependent on the direction we take the distance in, while Condition 3 is referred to as the triangular inequality. In some situations, it might still be useful to measure distance using a function which does not satisfy one or more of these conditions, however we would not refer to these as metric functions.

In the case of random variables and distributions, information theoretical distance notions are applied. Some of these are related to metrics, while others attempt to measure mutual or discriminatory information, or the extent to which one distribution can be used to infer about another. The foundation of information distance is entropy [Cover & Thomas, 2006], which is defined as

$$H(X) = -\sum_{x\in\chi} \Pr[X = x] \log \Pr[X = x]. \tag{1.4}$$

Where X is a discrete random variable, χ is the sample space of X. The logarithm is taken in base 2 (and can be used from here on), so we consider the entropy of a random variable X to be a measure of uncertainty in bits.

For two given flows with distributions $p(x)$ and $q(x)$, the relative entropy or Kullback-Leibler distance (KL distance) [Cover & Thomas, 2006] is defined as follows.

$$D(p,q) = \sum_{x\in\chi} p(x) \log \frac{p(x)}{q(x)} \tag{1.5}$$

where χ is the sample space of X.

It is straightforward to see from Eq. (1.5) that KL distance satisfies the first condition of a metric function, i.e. that $D(p,q) = 0 \Longleftrightarrow p = q$) and fails the second since it is not symmetric. It can be shown (e.g. see [Cover & Thomas, 2006]) that $D(p,q)$ is non-negative where p and q are distributions and we further note that it does not satisfy the triangular inequality.

There have been a few semi-metrics based on KL distance which correct for the problem of asymmetry, however still do not satisfy the triangular inequality. These include the Jeffrey distance and the Sibson distance.

The Jeffrey distance is given by,

$$D_J(p,q) = \frac{1}{2}\left[D(p,q) + D(q,p)\right]. \tag{1.6}$$

Rather than take the average of both directions, the Sibson distance averages the KL distance from $p(x)$ and $q(x)$ to their average over each x, i.e.

$$D_S(p,q) = \frac{1}{2}\left[D\left(p, \frac{1}{2}(p+q)\right) + D\left(q, \frac{1}{2}(p+q)\right)\right] \tag{1.7}$$

The Hellinger distance is closer to more traditional notions of distance, equivalent to the Euclidean metric with a root transformation of the variables [McLachlan, 1992]. It satisfies all the properties of a metric, and is defined as,

$$D_H(p,q) = \left[\sum_{x \in \chi} \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 \right]^{\frac{1}{2}}. \tag{1.8}$$

We can see that each of these distances will evaluate differences between distributions on a different scale. For instance, since $D_S(p,q)$ measures uses the relative entropy between each of the distributions and their average, we could expect it to consistently produce smaller values than D_J. For purposes of reliability analysis, what is more important is the ability to effectively discern between distinct types of distributions. We require the functions to be sensitive enough to identify key events, which could be flash crowds or attacks (and preferably be able to distinguish between them), but not so sensitive that fluctuations in line with normal flow patterns result in relatively large difference measures. Research has been conducted on the aforementioned similarity functions, comparing them in terms of such sensitivity assessments [Yu et al., 2009].

1.5 Performance Comparison for Information Distances

We can illustrate the differences between D_J, D_S and D_H in discerning between legitimate and attack traffic with the following numerical experiments. We consider Poisson and Normal distributions respectively, since Internet traffic is generally assumed to conform to these flow patterns. A combination of Normal distributions with varying parameters can be used to represent any type of data distribution, so the results here should give a reasonable indication of how each of the distance-type measures perform. In particular we are interested in 1) the sensitivity of each of the distances when comparing random distributions generated from the same parameters - i.e. the sensitivity to fluctuations that would occur naturally in the case of legitimate user behavior, and 2) the ability to detect and distinguish between distributions characteristic of abnormal events - i.e. malicious attacks and flash crowds.

First of all, two Normal flows ($\mu = 10, \delta = 1$) are arranged, and the three functions are used to measure the information distance. The simulation is conducted 100 times, and the results are shown in Figure 1.1.

We perform a similar evaluation using two generated Poisson flows with the value of $\lambda = 10$. The results are shown in Figure 1.2.

For two flows sharing the same distribution and parameter(s), the distance between them should ideally be close to 0. We note in Figure 1.1 and Figure 1.2, that Sibson's distance measure gives outputs on a scale much smaller than the Jeffrey and Hellinger distance.

Sensitivity to small changes in flow is another important factor in evaluating each measure's performance in monitoring base-rate traffic flows. We now investigate the

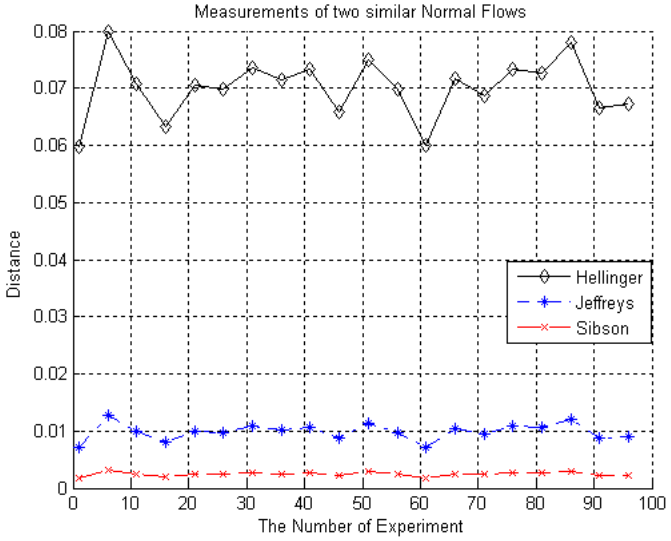

Fig. 1.1 Information distance for two Normal flows ($\mu = 10, \delta = 1$)under the different metrics.

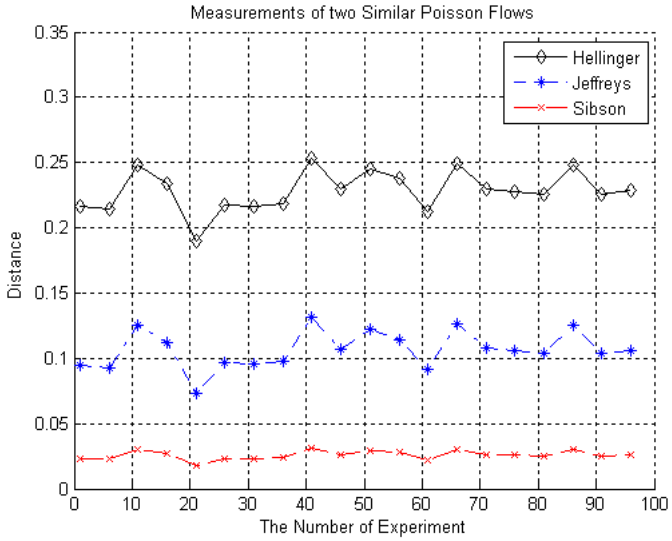

Fig. 1.2 Information distance for two Poisson flows $\lambda = 10$ under the different metrics.

behavior of each distance as the standard deviation of the Normal flow is adjusted from 0.1 to 3, i.e. 1% to 30% of the mean (keeping the value of $\mu = 10$). The results are shown in Figure 1.3.

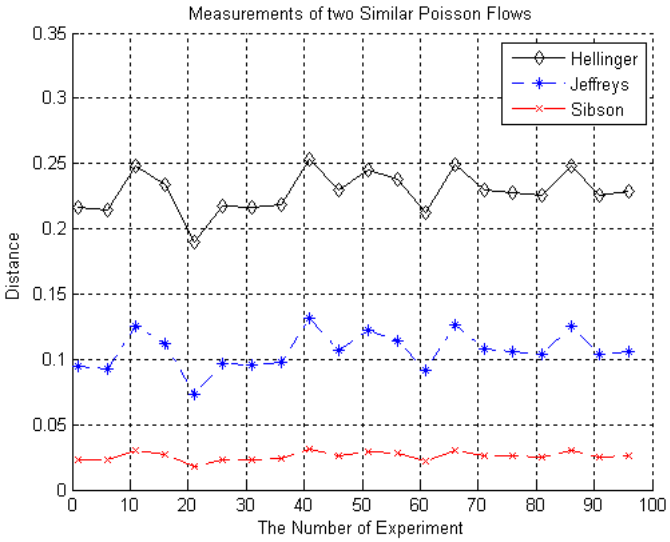

Fig. 1.3 The metric sensitivity of Normal flows ($\mu = 5$) against standard variations.

For the Poisson flows, we examine the sensitivity against arrival rate, which varies from 5 to 12. The results are shown in Figure 1.4.

Based on Figure 1.3 and 1.4, it is shown again that the Sibson's information distance remains quite low in terms of its output scale. The simulations demonstrated that it is quite stable where the parameters are gradually altered for both Normal and Poisson flows.

1.6 Summary

In this Chapter, we pointed out that a solid result of web data mining comes from a reliable data set, therefore, the data set for web data mining is critical for knowledge discovery. As the volume of web traffic is extraordinarily huge, we introduce the aggregation functions to deposit the traffic patterns with much smaller storage space while preserving the features of the original data. In order to filter out anomaly data, e.g. DDoS attack or botnet recruitment data, information distances are hired to carry out the task. Among the three information distances, we found that the Sibson's information distance is the best among them.

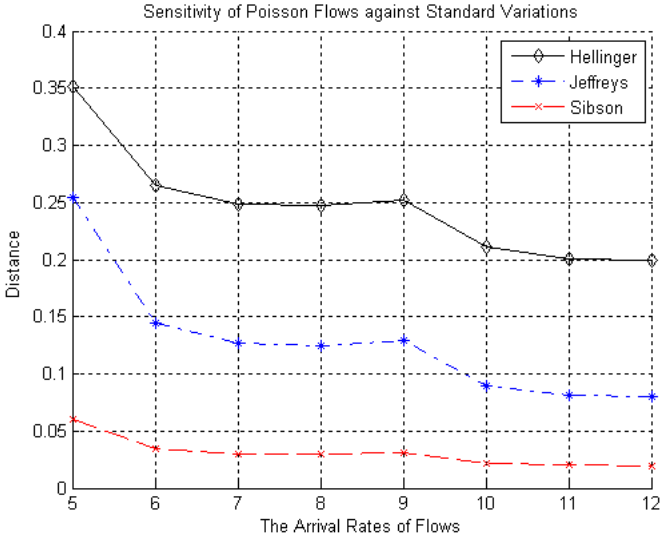

Fig. 1.4 The metric sensitivity of Poisson flows against arrival rate.

References

[Beliakov et al., 2007] Beliakov, G., A. Pradera, & T. Calvo 2007. Aggregation Functions: A Guide for Practitioners. Springer, Heidelberg, Berlin, New York.

[Cooley, 2000] Cooley, Robert Walker 2000. Web Usage Mining: Discovery and Application of Interestin Patterns from Web Data.

[Cover & Thomas, 2006] Cover, Thomas M., & Joy A. Thomas 2006. Elements of Information Theory. John Wiley & Sons.

[El-Atawy et al., 2009] El-Atawy, Adel, Ehab Al-Shaer, Tung Tran, & Raouf Boutaba 2009. Adaptive Early Packet Filtering for Protecting Firewalls against DoS Attacks. In Proceedings of the INFOCOM.

[Grabisch et al., 2009] Grabisch, M., J.-L. Marichal, R. Mesiar, & E. Pap 2009. Aggregation Functions. Cambridge University Press, Cambridge.

[Manavoglu et al., 2003] Manavoglu, Eren, Dmitry Pavlov, & C. Lee Giles 2003. Probabilistic User Behavior Models. Data Mining, IEEE International Conference on, 0:203.

[McLachlan, 1992] McLachlan, G J 1992. Discriminant analysis and statistical pattern recognition. Wiley-Interscience.

[Moore et al., 2006] Moore, David, Colleen Shannon, Douglas J. Brown, Geoffrey M. Voelker, & Stefan Savage 2006. Inferring Internet denial-of-service activity. ACM Transactions on Computer Systems, 24(2):115–139.

[Peng et al., 2007] Peng, Tao, Christopher Leckie, & Kotagiri Ramamohanarao 2007. Survey of network-based defense mechanisms countering the DoS and DDoS problems. ACM Computing Survey, 39(1).

[Srivastava et al., 2000] Srivastava, Jaideep, Robert Cooley, Mukund Deshpande, & Pang-Ning Tan 2000. Web usage mining: discovery and applications of usage patterns from Web data. SIGKDD Explor. Newsl., 1:12–23.

[Thing et al., 2007] Thing, Vrizlynn L. L., Morris Sloman, & Naranker Dulay 2007. A Survey of Bots Used for Distributed Denial of Service Attacks. In SEC, pages 229–240.

[Torra & Narukawa, 2007] Torra, V., & Y. Narukawa 2007. Modeling Decisions. Information Fusion and Aggregation Operators. Springer, Berlin, Heidelberg.

[Wang et al., 2007] Wang, Haining, Cheng Jin, & Kang G. Shin 2007. Defense against spoofed IP traffic using hop-count filtering. IEEE/ACM Transactions on Networking, 15(1):40–53.

[Yager, 1988] Yager, R.R. 1988. On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Transactions on Systems, Man and Cybernetics, 18:183–190.

[Yager & Filev, 1999] Yager, R.R., & D. P. Filev 1999. Induced ordered weighted averaging operators. IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics, 20(2):141–150.

[Yager & Beliakov, 2010] Yager, R. R., & G. Beliakov 2010. OWA operators in regression problems. IEEE Transactions on Fuzzy Systems, 18(1):106–113.

[Yu et al., 2008] Yu, Shui, Robin Doss, & Wanlei Zhou 2008. Information Theory Based Detection Against Network Behavior Mimicking DDoS Attacks. IEEE Communications Letters, 12(4):319–321.

[Yu et al., 2009] Yu, Shui, Theerasak Thapngam, Jianwen Liu, Su Wei, & Wanlei Zhou 2009. Discriminating DDoS Flows from Flash Crowds Using Information Distance. In Proceedings of the 3rd International Conference on Network and System Security, pages 351–356.