Electronics **2015**, *4*, 221-238; doi:10.3390/electronics4020221

OPEN ACCESS

electronics

ISSN 2079-9292

www.mdpi.com/journal/electronics

Article

Data-Throughput Enhancement Using Data Mining-Informed Cognitive Radio

Khashayar Kotobi [1,*], Philip B. Mainwaring [2], Conrad S. Tucker [3,4,5] and Sven G. Bilén [1,2,3]

[1] Department of Electrical Engineering, The Pennsylvania State University, University Park, PA 16802, USA; E-Mail: sbilen@psu.edu

[2] Department of Aerospace Engineering, The Pennsylvania State University, University Park, PA 16802, USA; E-Mail: pbm129@psu.edu

[3] School of Engineering Design, Technology, and Professional Programs, The Pennsylvania State University, University Park, PA 16802, USA; E-Mail: ctucker4@psu.edu

[4] Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA

[5] Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802, USA

* Author to whom correspondence should be addressed; E-Mail: kzk156@psu.edu; Tel.: +1-720-369-7255.

Academic Editors: Sanqing Hu, Lian Zhao and Nazanin Rahnavard

Received: 23 November 2014 / Accepted: 16 March 2015 / Published: 26 March 2015

Abstract: We propose the data mining-informed cognitive radio, which uses non-traditional data sources and data-mining techniques for decision making and improving the performance of a wireless network. To date, the application of information other than wireless channel data in cognitive radios has not been significantly studied. We use a novel dataset (Twitter traffic) as an indicator of network load in a wireless channel. Using this dataset, we present and test a series of predictive algorithms that show an improvement in wireless channel utilization over traditional collision-detection algorithms. Our results demonstrate the viability of using these novel datasets to inform and create more efficient cognitive radio networks.

Keywords: cognitive radio network; big data applications; data mining; wireless communication; data mining-informed cognitive radio

---

1. Introduction

1.1. Scarce Spectrum Problem

Demand is growing rapidly for wireless communication technologies, such as wireless data links, mobile telephones and wireless medical technologies. This demand places a significant burden on the limited wireless spectrum. One method to mitigate this lack of spectrum is to employ cognitive radio techniques in these wireless technologies. Prior research has explored cognitive radio aspects, such as spectrum access, truthful spectrum auctions [1–4] and dynamic spectrum rental. In this paper, we investigate how the performance of a cognitive radio network can be improved through the utilization of crowd-sourced social media information. In particular, we are interested in algorithms that can extract relevant information about the wireless channel condition and current and future user demands through sources of information that are not currently considered for this purpose.

1.2. Background on Cognitive Radio

The inefficient usage of the limited frequency spectrum makes it difficult to meet the increasing demand for wireless communication capacity [5–7]. Cognitive radio has been introduced as a solution to this problem. Cognitive radio is defined as "a radio that can change its transmitter parameters based on interaction with the environment in which it operates" [8]. The term "cognitive radio" was first introduced by Mitola in 1999 [9]. Cognitive radio is an evolved version of software-defined radio that can reconfigure itself, such that it can adapt its waveform parameters to the environment to meet higher-layer user demands for a high quality of service (e.g., voice over IP, video). Cognitive radio implementations fall between two extremes. At one end is the Mitola radio, a radio that collects information about all observable wireless information. As such, it is a theoretical construct that cannot be implemented in practice, but provides an ideal to aim for in cognitive radio research. At the other end is what can be practically implemented, which may be a spectrum-sensing cognitive radio with information about only the frequency spectrum [10]. In our work, we seek to move towards the Mitola radio ideal by employing more and varied information in our control loop, such that the cognitive radio can make more informed decisions.

One can define the cognitive ability of a radio as capturing and gathering information regarding the state of the environment, processing this information and then determining corrective action based on its findings. This cognitive process is not limited to monitoring the power level in a specific frequency band, but can also include the spatial and temporal variations in the radio environment due to the mobility and time dependency of most wireless devices. Based on the users' demands, these devices need to access a free and/or unused spectrum band at different times and/or locations. Reconfigurability empowers the radio to dynamically adapt to a changing radio environment [10]. This means that the cognitive radio will adjust to communicate in an appropriate frequency band and with a suitable waveform (*i.e.*, modulation type).

Spectral and temporal analysis of the radio spectrum reveals three broad categories of frequency band usage [11]:

- Frequency bands that are predominantly unoccupied;

- Frequency bands that are moderately occupied; and
- Frequency bands that are heavily occupied.

Note that a channel may be heavily occupied at one period in time, but not at another. Cognitive radio offers opportunistic usage of the frequency spectrum if permitted by primary users who currently "own" that slice of spectrum [12]. This process is called dynamic spectrum access, which may rely on algorithms and concepts found in game theory and network information theory.

A dynamic system is defined to be cognitive if it employs the perception-action cycle and has memory, attention and intelligence [13]. In the perception-action cycle depicted in Figure 1, a perceptor gathers measurements and sends them as feedback information to an actuator that uses this to control the perceptor via the environment. Memory is needed since the environment is nonstationary. The actuator prioritizes the allocation of limited resources, and feedback enables the presence of intelligence in this system by providing the perception of the environment to the actuator.

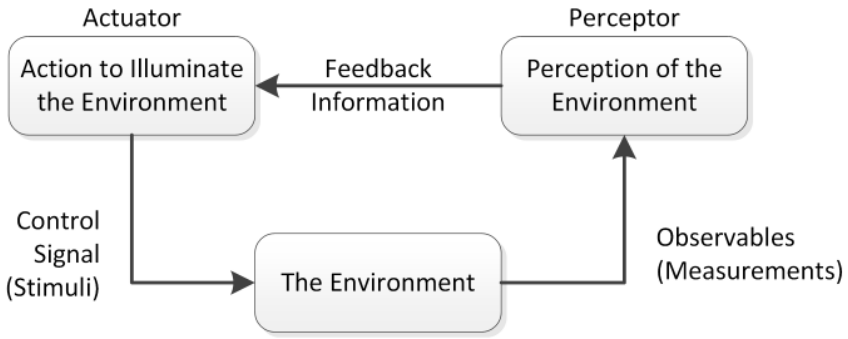

Figure 1. The perception-action cycle of a cognitive dynamic system in its most generic sense. Reprinted with permission from Cambridge, 2012 [13].

Three fundamental cognitive radio tasks based on the perception-action cycle are introduced in [11]. For the receiver, information gathering and analysis must be performed to determine the condition of the dynamic radio environment. For the transmitter, power budgeting and dynamic spectrum access based on information regarding the presence of the primary users must be calculated and executed. Finally, there must be a feedback channel between the transmitter and receiver regarding information about the radio environment. A cognitive radio can more effectively adapt to the radio environment if it can cooperate with other cognitive radios [14].

A cognitive radio network seeks to serve the individual communication requirements of multiple primary and secondary users. In doing so, three practical challenges arise [13]. First, the vacancies in spectrum come and go due to temporary usage of the spectrum by a licensed primary user. Finding these vacancies can be accomplished more efficiently in a cooperative manner by the secondary users in a cognitive radio network. Second, for each cognitive radio, information gathered by the receiver components must be processed and sent to the transmitter side. This will induce a delay in the feedback channel. Third, the security of the cognitive radio network can be compromised by malicious users in different locations and time frames. These practical issues need to be solved; in addition, solutions that use game theory to promote a cooperative strategy between the secondary users are also necessary.

1.3. Big Data Framework

As mentioned above, cognitive radio is being explored to address the scarce spectrum problem. The medium access control (MAC) protocol for any realizable system allows wireless users to use the frequency channels based on the current state of the network. For example, in [15], various MAC protocols are investigated for use in cognitive radio based on different methods for users to start communication. In this article, the usage of data-mining techniques in MAC protocols for cognitive radio is studied.

The term "big data", often used when referring to data mining, extends beyond the volume of data acquired to include also the velocity (how fast the data are being transmitted), the variety (the different data types included), veracity (the accuracy or truthfulness of the data) and value (the tangible benefits that the data provide). The authors have demonstrated that large-scale social media networks exhibit the five Vs of "big data" and can serve as a viable source of real-time knowledge extraction though data mining [16,17].

Using techniques, such as data mining, and including them in the perception-action cycle of a cognitive radio is an emerging field of study. Recently, a vision was presented for the use of "big data" techniques to enhance the performance of a cognitive radio network [18]. The big data vision is foundational for illustrating cognitive networked sensing, cognitive radar, smart grid and cognitive radio networks. The data employed in [18], however, are only those concerning the wireless channel, which are then employed to enhance the decision-making process regarding the usage of available wireless channels.

1.4. Main Contributions

The application of data sources other than wireless channel data in cognitive radio has not been significantly studied. These new information sets can be used to predict the traffic, channel condition and other conditions of a wireless network. As an example, we demonstrate herein that data extracted from a social media network by means of data mining can inform a cognitive radio.

Cognitive radios work by collecting information about a statistically varying environment and then applying methods and algorithms that react to this collected information to maximize certain performance goals. In this work, we employ data mining and game-theoretic techniques that employ new environmental data. Specifically, we investigate a new cognitive radio scheme that uses crowd-sourced social media information obtained through data mining of a large-scale social media network. We employ game theoretic algorithms for adaptation and reaction to a varying radio environment. The performance improvements from these adaptations help demonstrate the merit of what we call the "data mining-informed cognitive radio".

1.5. Proposed Cognitive Radio

The research framework that we employ to develop more advanced cognitive radio networks consists of two thrusts. First, we use sources of information other than wireless channel information and collect relevant information using data-mining techniques to inform cognitive radio networks. The main goal

of this thrust is to understand these novel sources of information and to find new methods to collect and analyze data that are not directly correlated to the wireless channel information, but which are relevant to channel usage. The second thrust employs appropriate game theoretic techniques to better utilize the spectrum and to perform resource sharing between the secondary wireless users in a wireless network. In this paper, we introduce smarter cognitive radio nodes and networks by way of improved algorithms, and it is shown that, based on crowd-sourced information, one can increase data transmission throughput.

Wireless channel information is currently the primary feedback information source used in cognitive radio. This information helps the various radio network nodes to adapt their waveforms, e.g., frequency, data transmission rate, modulation, *etc*. If the channel is used by other primary or secondary users, the user can detect the channel load and back off. However, there are other sources of information that may be helpful in predicting the near-term or future conditions of a wireless channel. For example, information about nearby weather conditions or forecasting a rainy day will help the coordinator and/or secondary users to use more robust modulations compared to more data-rate-efficient modulations. Predicting a sunny day would imply using more data-rate-efficient modulations. Another example is the acquisition of information about an emergency situation, such as a fire at a school, by data mining of social media and other sources. In this case, the moderator or secondary users must prioritize data transmission related to the emergency event over their own data. Gathering these new data will add additional information for better decision making and, thus, improve the performance of a wireless network.

1.6. Paper Organization

The reminder of the paper is organized as follows. In Section 2, we consider general cognitive radio concepts, investigate their current limitations and introduce methods for improvement in performance regarding spectrum usage. In Section 3, the proposed algorithms to improve channel traffic allocation are introduced. Simulation results are presented in Section 4. In Section 5, we summarize our results and present conclusions based on the results presented in this paper.

2. Problem Definition

2.1. Big Data Framework for Wireless Communication

As mentioned in Section 1.3, wireless channel information is not the only relevant information that can be considered for a cognitive radio or by a cognitive radio network administrator. In this section, an example of a data source that is not directly related to the wireless channel information is presented, and several algorithms to use these data are presented with the goal of improving the throughput of current cognitive radio networks.

2.2. Problem Definition

With the goal of enhancing the performance of a cognitive radio network, there has been limited study on the use of information obtained via data mining of datasets that are correlated or uncorrelated

with wireless channel information, only the use of wireless channel information itself [18]. We posit, however, that additional information can lead to data traffic prediction, which can be used to enhance the performance of cognitive radio. Better understanding of the wireless channel capacity based on the wireless channel information is critical for finding the best transmission plan, modulation and rate, although one can argue there are other sources of data available that can help predict the near or far future environment of the network. For example, one can predict data traffic by using information gathered on the daily schedule of events in a restaurant. A popular concert in the city center of a European town can cause collision and denial of service in wireless mobile phone communication. Using this information and allocating more frequency spectrum to use in those cells in which the event is taking place will help avoid congestion in the network.

The emergence of low-cost mobile communication devices and digital storage technologies is enabling the rapid creation and dissemination of information on a global scale. Digital information, ranging from user-generated data (e.g., captured through social media, such as Twitter, Facebook, Google +, *etc*.), to data generated through industry and government efforts, is establishing a new dimension of social-driven knowledge discovery. These data, fed through the appropriate cognitive engines, might inform cognitive radio beyond what radio parameters alone can do and, in the process, expand bandwidth. The challenge facing us is not the lack of digital information, but rather the synthesis of large-scale, multi-domain data and their transformation into actionable information by cognitive radio systems. Figure 2 shows how data may come from cognitive radios and/or other sources. Some of the cognitive radios may be collocated in a geographical region (e.g., coffee shop, mall, concert venue, *etc*.) that also has data associated with it, as well as data generated by the cognitive radios in that region, all of which may be mined [19].

2.3. Wireless Channel Modeling

Knowledge of wireless channel traffic is one of the parameters that can be used to improve the performance of the cognitive radio network. Conventionally, this might be modeled by channel signal-to-noise ratio (SNR). If the SNR in the receiver is higher than a specific threshold, one can assume that the wireless channel is occupied. However, channel occupying level might also be modeled by the number of Tweets or Facebook posts per unit time in a specific location (a more nuanced analysis might also include other aspects of the social media data obtained through the five Vs discussed in Section 1.3). We demonstrate here a collision-free data transmission scheme that mimics the channel capacity by using the current and past channel traffic and the channel information, which will result in a close-to-optimum throughput. For this demonstration, we use the dataset of received signal strength (RSS) measurements for WiFi signals collected at the University of Colorado as our wireless channel information (obtained from CRAWDAD: Community Resource for Archiving Wireless Data At Dartmouth at http://crawdad.cs.dartmouth.edu/). We use the number of Tweets (actual data for a specific location in New York City) as an indicator of demand for data transmission over the wireless network. The Tweets used in the simulations have geolocation tags, which are used to extract the data traffic for a specific location.

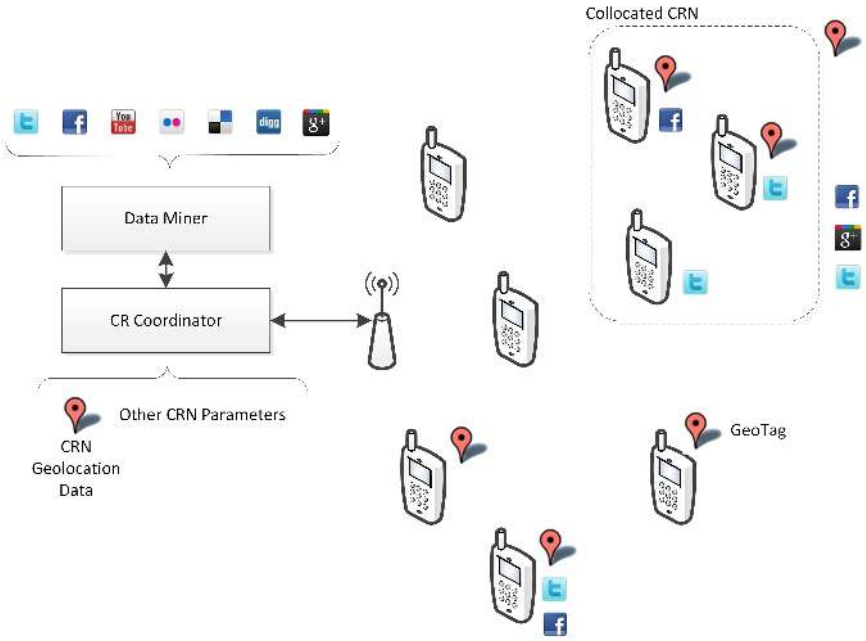

Figure 2. The proposed data mining-informed cognitive radio (DMICR) system utilizes other data sources in addition to the usual parameters employed in cognitive radio networks [19].

To simulate a wireless network environment, our algorithm uses wireless channel characteristics measured for a particular channel (measured at the University of Colorado). We use Rayleigh fading to model the multipath fading of the wireless channels. We then impose a set of unrelated network loads (here, in the form of Twitter data from the New York City area) to measure the effectiveness of the algorithm. By employing uncorrelated data, we avoid the scenario in which network traffic for a wireless network strongly matches its signal strength, *i.e.*, a poor-quality wireless signal discourages its use. In this simulation setup, we use Twitter traffic (more specifically, the publication of Tweets as opposed to reading of Twitter feeds) as an indicator of the actual network load. Various studies (e.g., [20,21]) show that a certain fraction of a wireless network load is Twitter traffic. Although that fraction may depend on the time and location, for this study, we assume that it is a fixed fraction.

The actual wireless channel capacity depends on many factors, including the coding scheme used in the physical layer, but here, we use the Shannon capacity to correlate the RSS measurements and the channel's available capacity. Based on RSS measurements, the channel Shannon capacity can be seen in Figure 3 for 4,000 min with one-minute resolution. For the demonstration of the algorithms proposed, we plot only the first 120 min, such that the channel variations can be observed. The channel usage and capacity are normalized by the maximum channel capacity, and their variations over time are plotted in Figure 4. As shown, the demand for usage can be higher than the channel capacity during some time periods, which results in collisions in the current medium-access algorithms. During other time slots, the wireless channel is underutilized. In order to achieve a closer-to-optimum throughput, the radio network should delay transmitting data packets in an over-utilized channel until the wireless channel is underutilized. This requires that the layers higher than the physical layer need to have understanding of the physical layer for wireless communication.

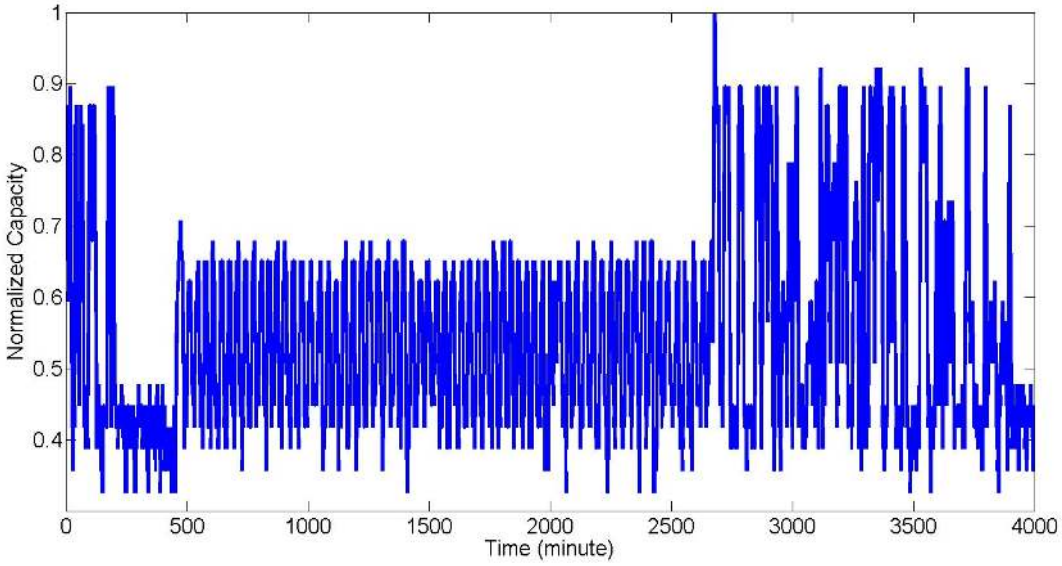

Figure 3. Variation in calculated point-to-point communication Shannon capacity of a WiFi network based on the received signal strength measurements as function of the signal-to-noise ratio with respect to time.

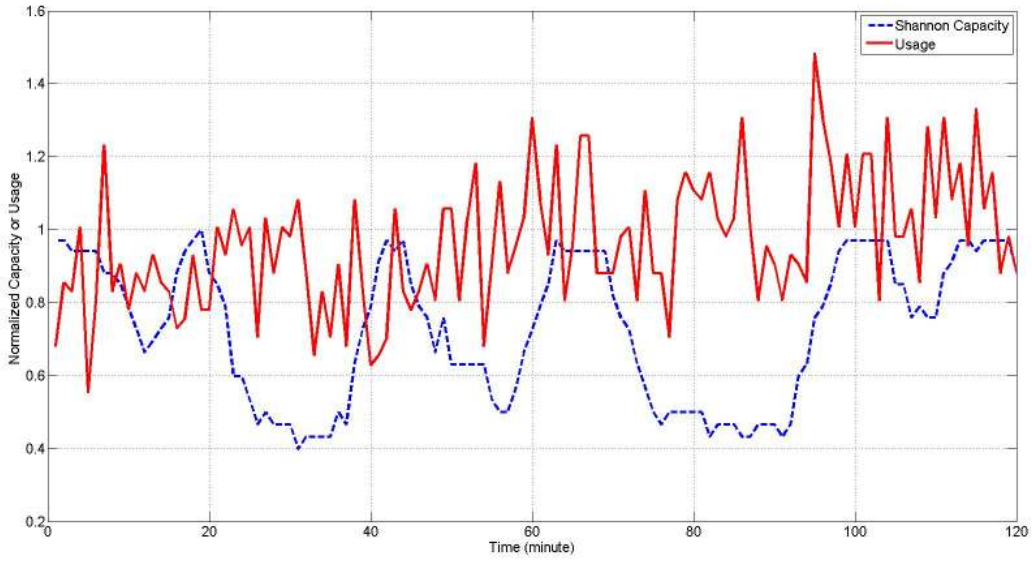

Figure 4. Comparison between a given demand for data transmission as a function of the number of the Tweets and calculated point-to-point communication Shannon capacity of a WiFi network based on the received signal strength measurements as function of the signal-to-noise ratio with respect to time.

3. Proposed Algorithms

3.1. Data Transmission Algorithms

In the conventional implementation of channel access, data packets are dropped in the case of a collision, which results in a throughput far from the channel capacity, as shown in Figure 5. This process is presented in Method 1.

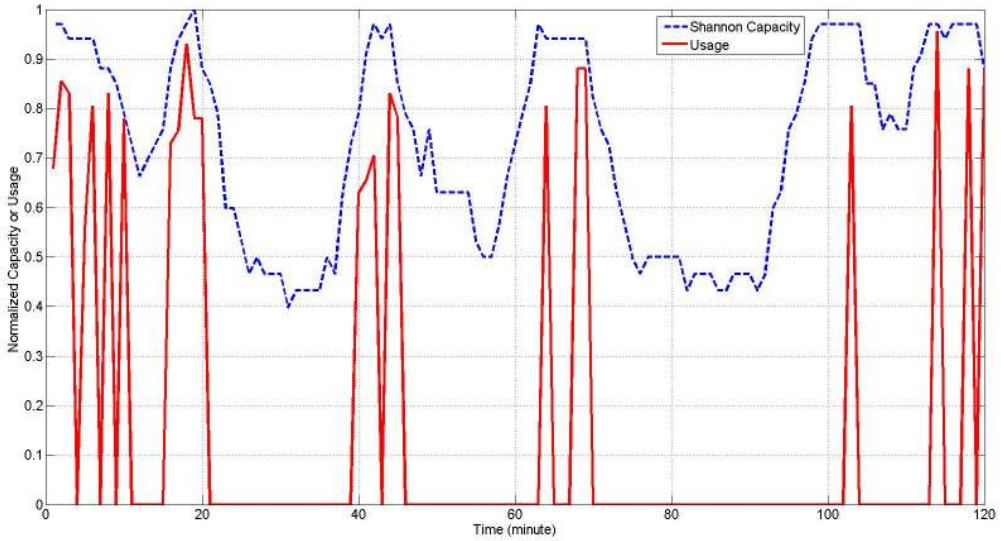

Figure 5. Comparison between normalized channel usage with the dropping of data packets in the event of collision and the calculated point-to-point communication Shannon capacity of a wireless channel with respect to time.

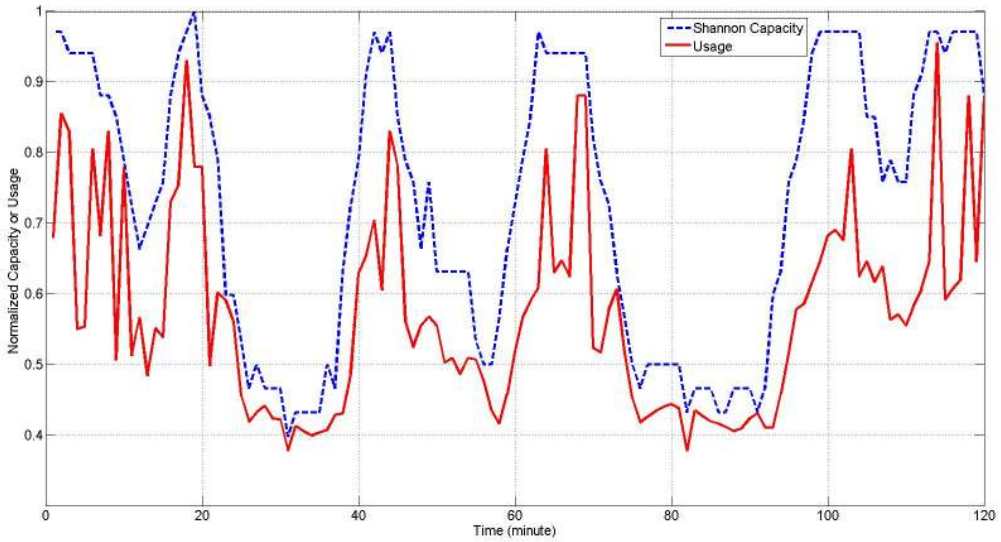

Figure 6. Comparison between the normalized channel usage with channel capacity prediction based on the algorithm in Method 3 with $N = 3$ and capacity with respect to time.

Method 1 Algorithm for data throughput with collisions.

for each t in $\{1, 2, ..., M\}$ **do** Find all capacity in $\{1, 2, ..., t\}$
  if Capacity(t) $<$ Demand(t) **then**
    Data$(t) = 0$

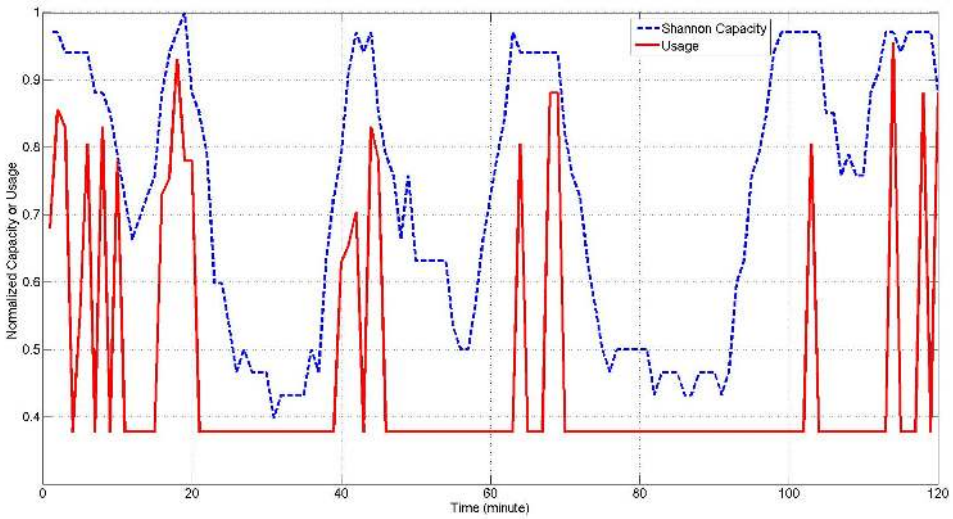

Figure 7. Comparison between normalized channel usage with using the dropping of non-real-time data packets in the event of collision method and calculated point-to-point communication Shannon capacity with respect to time.

If we categorize data packets into two data types, one for real-time applications, like voice over IP or video broadcasting, and the other as non-real-time data, like downloading a file, then we can choose to transmit only the real-time data types when we predict that the collisions will occur, which is based on the current change in the data usage and the current channel capacity. The algorithm for this is presented in Method 2. The implementation of this method is depicted in Figure 7, wherein it can be seen that the throughput is increased compared to that of the former method. The throughput is the integral of usage over time.

Method 2 Algorithm for prioritizing real-time data.

for each t in $\{1, 2, ..., M\}$ **do** Find all capacity in $\{1, 2, ..., t\}$
  if Capacity(t) $<$ Demand(t) **then**
    Data(t) $=$ Data$_{\text{Real}}$
    if Capacity(t) $<$ Data(t) **then**
      Data(t) $=$ Data$_{\text{Real}}$

To improve the throughput of the cognitive radio network, we can shape the transmission of the non-real-time packets based on the prediction of the channel capacity. Our prediction method uses the assumption that the channel capacity will change at the rate at which it was changing one or more time slots previously. In the case that the prediction is not true, the non-real-time data will be lost due to a collision according to Method 3. The results of this proposed method are shown in Figure 6 when

using the last three time slots, *i.e.*, $N = 3$. The throughput is significantly higher than Method 2, shown in Figure 7.

3.2. Incorporating Network Burstiness

Traditional Ethernet traffic exhibits a long-tailed probability distribution rather than a Poisson distribution, because of its self-similarity. Wireless networks have similarly shown this property [22]. This means that predicting Ethernet or wireless traffic is difficult: the self-similar nature of the traffic indicates that there is no defining length characteristic of network traffic bursts. This self-similarity increases as the number of network traffic sources increases [23].

Incorporating a linear weighting element of "burstiness" to our traffic prediction algorithm can help ameliorate issues with the Poisson distribution model. Based on the time fidelity of the data, an interval of prior traffic can be analyzed for a proportional burst factor. This factor can then be used to linearly weight the prediction of available capacity. There exists a wide variety of traffic prediction techniques that incorporate the self-similar behavior of network traffic: neural networks, auto-regressive integrated moving average models and alpha-stable models [24–26]. These models are a trade off between traffic prediction accuracy and the complexity of implementation, as well as the cost of performing such analysis. We choose a simple linear weighting to demonstrate the utility of incorporating self-similarity in network traffic prediction for many-user systems.

To analyze the burst factor, the number, size and length of bursts are computed by the method of [27] and given in a form modified for $N = 3$ in Method 4. A burst is defined as a period during which the traffic exceeds the average traffic of the network (over some tolerance factor of time). Once these burst characteristics are calculated, the network traffic can be characterized as either bursty or non-bursty (*i.e.*, steady). The analysis of these network traffic data show them to be primarily bursty with a low average load.

Having determined periods of network traffic that lie within bursts, the algorithm can either (1) send additional traffic during burst periods or (2) send additional traffic during steady, non-burst periods. We now develop a modified version of Method 3 with $N = 3$ to include a weighting factor proportionate to burst or steady periods.

Method 3 Algorithm for predicting channel condition and demand for N previous time slots.

for each t in $\{1, 2, ..., M\}$ **do** Find all capacity in $\{1, 2, ..., t\}$

  if $\text{Capacity}(t) < \text{Demand}(t)$ **then**

    $\text{Data}(t) = \text{Data}_{\text{Real}} + \text{Capacity}(t-1) - \text{Demand}(t-1) + \text{Capacity}(t-2)$

    $- \text{Demand}(t-2) + ... + \text{Capacity}(t-N) - \text{Demand}(t-N)$

  if $\text{Capacity}(t) < \text{Data}(t)$ **then**

    $\text{Data}(t) = \text{Data}_{\text{Real}}$

Method 4 Finding bursts in a time period.

for each t in $\{1, 2, ..., M\}$ **do** Find all capacity in $\{1, 2, ..., t\}$
  if Data(t) $>$ AverageData **then**
    burstStart $= t$
    burstData $=$ Data(t)
    for each tPrime in $\{t, t+1, t+2, ..., M\}$ **do**
      if Data(tPrime) $<$ AverageData **then**
        if Data(tPrime $- t$) $>$ ToleranceFactor **then**
          burstEnd $=$ tPrime $- 1$
          burstsPerTime$[t]$ $=$ Burst(burstStart, burstEnd, burstData)
        else
          burstData $=$ burstData $+$ Data(tPrime)
          continue

Method 5 Algorithm for predicting channel condition and demand incorporating traffic bursts.

for each t in $\{1, 2, ..., M\}$ **do** Find all capacity in $\{1, 2, ..., t\}$
  if Capacity(t) $<$ Demand(t) **then**
    PredictiveCapacity $=$ DataReal $+$ Capacity$(t-1)$ $-$ Demand$(t-1)$ $+$ Capacity$(t-2)$
    $-$Demand$(t-2)$ $+$ Capacity$(t-3)$ $-$ Demand$(t-3)$
    BurstProportionCapacity $=$ DataReal $+$ (BurstLength(t)/TimePeriod) $\times$ Demand(t)
    NonBurstProportionCapacity $=$ DataReal $+$ (TimePeriod $-$ (BurstLength(t)/TimePeriod))
    $\times$Demand(t)
  MaximumPrediction $=$ DataReal
for Prediction in PredictiveCapacity, BurstProportionCapacity, NonBurstProportionCapacity **do**
  if Prediction $>$ MaximumPrediction **and** Prediction $<$ Capacity(t) **then**
    MaximumPrediction $=$ Prediction
  Data(t) $=$ MaximumPrediction

This revised algorithm (presented in Method 5) incorporates burst characteristics to find an improved data transmission rate. It considers three factors: the network load steadiness (as a proportion of time), the network load burstiness (as a proportion of time) and the prediction for the network load (*i.e.*, Method 3 with $N = 3$). Using the time proportions for burst traffic and steady traffic, the algorithm generates a weighted data rate for both scenarios. It compares the three possible data rates and uses the method that results in the maximum usage of the channel without exceeding the capacity. We show here the results for both approaches (Figure 8 for burst, Figure 9 for steady) for one traffic dataset. What these results show is that the inclusion of burstiness results in slight throughput enhancement, but the additional computational complexity may not warrant its use.

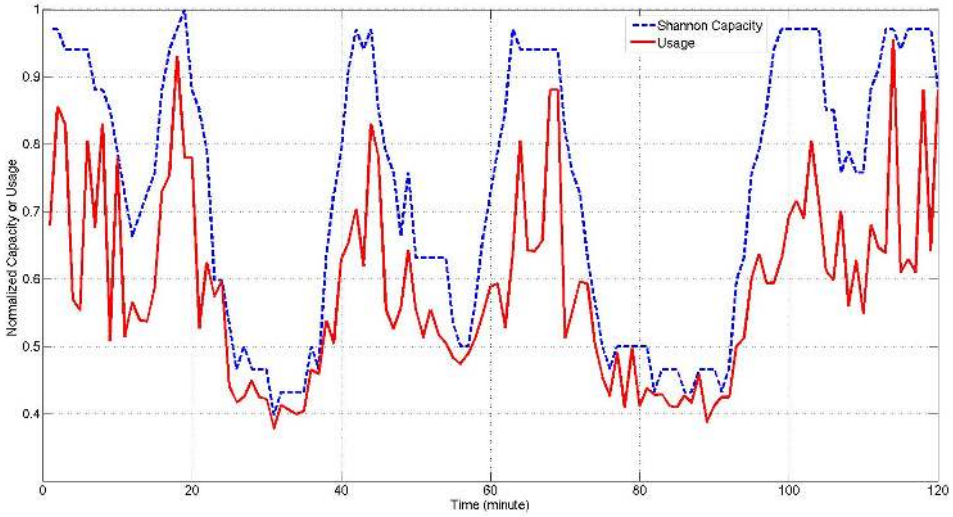

Figure 8. Comparison between normalized channel usage with channel capacity prediction based on a three-point prediction algorithm incorporating data transmission during burst traffic.

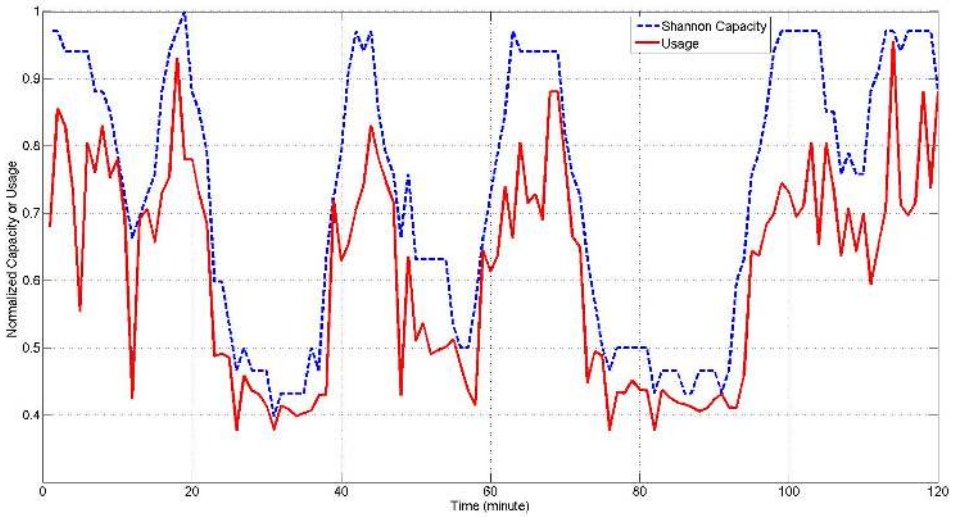

Figure 9. Comparison between normalized channel usage with channel capacity prediction based on a three-point prediction algorithm incorporating data transmission during non-bursty (steady) traffic period.

4. Results and Discussion

4.1. Simulation Setup

In this section, we present the simulation setup and the datasets used to investigate the proposed algorithms for improving the throughput of the wireless network. Here, we implement our methods to improve the throughput of the MAC protocol used for different network loads.

4.2. Throughput

One can compare the throughput when using Method 3 for different values of N. For the simulation environment we have developed, we have determined that the optimal throughput is achieved when $N = 3$; however, in actuality, minimal throughput gain is found beyond $N = 2$. To visualize this, the results for three different network loads are shown in Figure 10, which shows throughput with $N = 1$ to be much higher than with $N = 0$ and $N = 2$ and $N = 3$, achieving some small increase over $N = 1$. Given the extra computational complexity for higher values of N, we suggest using $N = 2$. Other network environments may find a different optimal value.

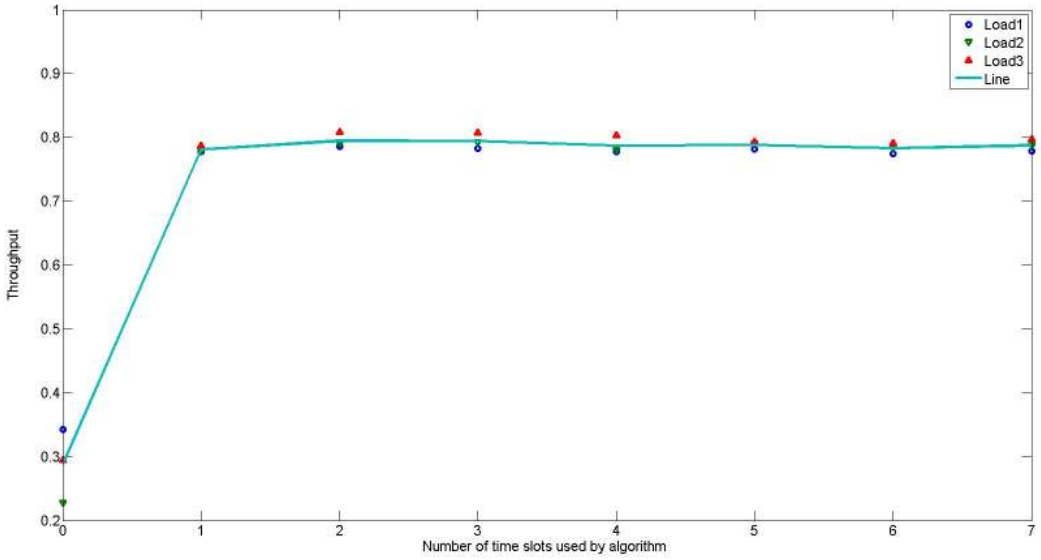

Figure 10. The channel throughput *versus* the number of time slots N used by Method 3 for three different network load levels.

In order to assess the performance of our network protocol, we compare it against a standard protocol, such as slotted Aloha, which is equivalent to $N = 0$ in our protocol. Figure 10 shows the throughput of our protocol under three different network loads for increasing values of N. We see that the throughput of our network is similar to that of slotted Aloha for $N = 0$, but much larger when $N \geq 1$. Throughput improvements for $N \geq 1$ are due to predicting the network traffic and shaping the data traffic. As similar patterns of improvement are seen under all three network loads, this demonstrates the generality of our proposed algorithms.

This work employed a dataset built from several sources, as discussed in Section 2.3. The correlation between the number of Tweets sampled in a time slot and the network traffic needs can be used to further fine tune the throughput achieved and needs to be investigated in future work. In future work, we will capture and use the actual network traffic to feed our algorithms.

4.3. Simulation Methods

We used MATLAB 2013 to implement the algorithms. C# and Visual Studio 2012 were used to compute the list of bursts for network traffic for a given time period, while the predictive algorithms were implemented using MATLAB.

Data existing in large-scale social networks, such as Twitter, can be acquired through application programming interfaces (APIs) provided by these systems. In the case of Twitter, their API can be accessed by visiting https://dev.twitter.com/. The API handles queries made about users of the social network, such as: (i) time of a post; (ii) location (if available); (iii) message content; *etc*. [28]. Using a vector of keywords (e.g., find all Tweets containing the words ([restaurant, friends, pictures)), messages pertaining to a given query are returned and subsequently can be stored in a database. Data mining techniques can then be employed on these data.

4.4. Simulation Supplementary Materials

The dataset of received signal strength (RSS) measurements for WiFi signals was collected at the University of Colorado and obtained from CRAWDAD, http://crawdad.cs.dartmouth.edu/.

The number of Tweets (actual data for a specific location in New York City) with geolocation tags was obtained from dataset available to Conrad S. Tucker's research group.

5. Summary and Conclusions

Traditional data sources for cognitive radio networks generally only include the properties of the wireless channel. More efficient cognitive radio networks can be constructed by the analysis of additional relevant data sources (among them, social media, like Twitter). By applying these new data sources, we can build predictive algorithms for the network load of the wireless network that increase the network throughput. We call such a radio a data mining-informed cognitive radio.

For our simulations, we constructed a test dataset using Twitter traffic as a model of network load in a measured wireless channel. We developed new algorithms that employ both predictive techniques, as well as network traffic analysis. By measuring their performance against traditional collision detection algorithms, we have shown that these algorithms improve the utilization (*i.e.*, data throughput) of the wireless channel.

In future work, additional improvements to cognitive radio networks can be made by the introduction of game theoretic techniques. Altruistic cognitive players can be introduced to monitor and police the network. The combination of further data mining and game theory will increase the performance of cognitive radio networks.

Acknowledgments

The authors thank Suppawong Tuarob for his contribution in providing the dataset of Tweets.

Author Contributions

Sven G. Bilén formulated the concept of the data mining-informed cognitive radio; Khashayar Kotobi and Sven G. Bilén contributed to the problem formulation and background material; Khashayar Kotobi, Philip B. Mainwaring and Sven G. Bilén contributed to the algorithm simulations and the interpretation of the results; Conrad S. Tucker supervised the extraction of the data-mined Tweets; All authors were involved in the manuscript preparation. All authors approved the final manuscript.

Conflicts of Interest

The authors would like to declare that there are no conflicts of interests.

References

1. Sodagari, S.; Attar, A.; Bilén, S.G. Strategies to achieve truthful spectrum auctions for cognitive radio networks based on mechanism design. In Proceedings of the 2010 IEEE Symposium on New Frontiers in Dynamic Spectrum, Singapore, 6–9 April 2010; pp. 1–6.
2. Sodagari, S.; Attar, A.; Leung V.C.M.; Bilén, S.G. Time-optimized and truthful dynamic spectrum rental mechanism. In Proceedings of the 72nd IEEE Vehicular Technology Conference Fall (VTC 2010-Fall), Ottawa, ON, Canada, 6–9 September 2010; pp. 1–5.
3. Jana, S.; Zeng, K.; Cheng, W.; Mohapatra, P. Trusted Collaborative Spectrum Sensing for Mobile Cognitive Radio Networks. *Inf. Forensics Secur.* **2013**, *8*, 1497–1507.
4. Fatemieh, O.; Chandra, R.; Gunter, C.A. Secure collaborative sensing for crowdsourcing spectrum data in white space networks. In Proceedings of the 2010 IEEE Symposium on New Frontiers in Dynamic Spectrum, Singapore, 6–9 April 2010 pp. 1–12.
5. Akyildiz, I.F.; Altunbasak, Y.; Fekri, F.; Sivakumar, R. AdaptNet: Adaptive protocol suite for next generation wireless internet. *IEEE Commun. Mag.* **2004**, *42*, 128–138.
6. Giannoulis, A.; Patras, P.; Knightly, E.W. *Mobile Access of Wide-Spectrum Networks: Design, Deployment and Experimental Evaluation*; 2012, arXiv:1204.4847. arXiv.org e-Print archive. Available online: http://arxiv.org/abs/physics/0402096 (accessed on 21 April 2012).
7. Wang, B.; Wu, Y.; Liu, K.J. Game theory for cognitive radio networks: An overview. *Comput. Netw.* **2010**, *54*, 2537–2561.
8. FCC. Notice of proposed rule making and order. ET Docket No 03-222, December 2003; pp. 1–21.
9. Mitola, J.; Maguire, G.Q., Jr. Cognitive radio: Making software radios more personal. *IEEE Personal Commun.* **1999**, *6*, 13–18.
10. Akyildiz, I.F.; Lee, W.; Vuran, M.C.; Mohanty, S. Next generation dynamic spectrum access cognitive radio wireless networks: A survey. *Comput. Netw.* **2006**, *50*, 2127–2159.
11. Haykin, S. Cognitive radio: Brain-empowered wireless communications. *IEEE J. Sel. Areas Commun.* **2005**, *23*, 201–220.
12. Gyucek, T.; Huseyin, A. A survey of spectrum sensing algorithms for cognitive radio applications. *Commun. Surv. Tutor. IEEE* **2009**, *11*, 116–130.

13. Haykin, S. *Cognitive Dynamic Systems Perception Action Cycle, Radar, and Radio*; Cambridge University Press: New York, NY, USA, 2012.
14. Jondral, F.K. Software-defined radio basics and evolution to cognitive radio. *EURASIP J. Wirel. Commun. Netw.* **2005**, *3*, 275–283.
15. Hu, S.; Yao, Y.; Yang, Z. MAC protocol identification using support vector machines for cognitive radio networks. *IEEE Wirel. Commun.* **2014**, *21*, 52–60.
16. Bodnar, T.; Tucker, C.S.; Hopkinson, K.; Bilén, S.G. Increasing the veracity of event detection on social media networks through user trust modeling. In Proceedings of the 2014 IEEE International Conference on Big Data, Washington, DC, USA, 27–30 October 2014.
17. Yin, P.; Ram, N.; Lee, W.; Tucker, C.S.; Khandelwal, S.; Salathe, M. Two sides of a coin: Separating personal communication and public dissemination accounts in Twitter. In Proceedings of the 2014 Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Tainan, Taiwan, 13–16 May 2014.
18. Qiu, R.; Wicks, M. *Cognitive Networked Sensing and Big Data*; Springer: New York, NY, USA, 2014.
19. Bilén, S.G.; Kotobi, K.; Tucker, C.S. Data mining–informed cognitive radio networks. In Proceedings of the 2014 New England Workshop on Software Defined Radio (NEWSDR 14), Boston, MA, USA, 6 June 2014.
20. Li, Y.; Zhou, G.; Ruddy, G.; Cutler, B. A measurement-based prioritization scheme for smartphone applications. *Wirel. Personal Commun.* **2014**, *78*, 333–346.
21. Das, A.K.; Pathak, P.H.; Chuah, C.N.; Mohapatra, P. Contextual localization through network traffic analysis. In Proceedings of the IEEE INFOCOM, Toronto, ON, Canada, 27 April–2 May 2014.
22. Basgeet, D.R.; Irvine, J.; Munro, A.; Dugenie, P.; Kaleshi, D.; Lazaro, O. Impact of mobility on aggregate traffic in mobile multimedia system. In Proceedings of the 5th International Symposium on Wireless Personal Multimedia Communications, Honolulu, HI, USA, 27–30 October 2002.
23. Leland, W.E.; Taqqu, M.S.; Willinger, W.; Wilson, D.V. On the self-similar nature of Ethernet traffic. *ACM SIGCOMM Comput. Commun. Rev.* **1993**, *23*, 183–193.
24. Zhou, B.; He, D.; Sun, Z.; Ng, W.H. Network traffic modeling and prediction with ARIMA/GARCH. In Proceeding of HET-NETs Conference, Ilkley, West Yorkshire, UK, 18–20 July 2005.
25. Sang, A.; San-qi L. A predictability analysis of network traffic. INFOCOM 2000. In Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies, Tel Aviv, Israel, 26–30 March 2000.
26. Ge, X.; Shaokai, Y.; Won-Sik, Y.; Yong-Deak, K. A new prediction method of alpha-stable processes for self- imilar traffic. In Proceedings of the 2004 Global Telecommunications Conference GLOBECOM '04 IEEE Dallas, TX, USA, 29 November–3 December 2004.
27. Krzanowski, R. Burst of packets and burstiness. In Proceedings of the 66th IETF meeting, Quebec, QC, Canada, 9–14 July 2006.

28. Russell, M.A. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2013.