Network Resilience: Exploring Cascading Failures within BGP*

E.G. Coffman Jr.
Dept. of Electrical Engineering
Columbia University
New York, NY 10027
egc@ee.columbia.edu

Zihui Ge
Dept. of Computer Science
University of Massachusetts
Amherst, MA 01003
gezihui@cs.umass.edu

Vishal Misra†
Dept. of Computer Science
Columbia University
New York, NY 10027
misra@cs.columbia.edu

Don Towsley
Dept. of Computer Science
University of Massachusetts
Amherst, MA 01003
towsley@cs.umass.edu

Abstract

Recent studies [1] have revealed vulnerabilities in the routing infrastructure of the Internet. It has been conjectured that these vulnerabilities could lead to cascading failures. In this paper we develop simple models for the interaction of routers, looking specifically at the clique topology. We construct two related models, and our analysis indicates that it is indeed possible to have cascading failures in systems such as a BGP clique. We encounter phase transitions in both our models, and we explore the dependence of system parameters on the nature and intensity of the phase transitions. Finally, we comment on the insights that we gain from our analysis.

1 Introduction

The Internet is a large collection of Autonomous Systems (AS). There are various glues that hold together this massive system, but the most important of these is the routing infrastructure. The Internet routing protocols maintain connectivity between and within AS's, and are designed to automatically reconfigure and recompute routing tables when they detect a link failure. This computation starts locally around the failure point, and then the information propagates through the Internet. While telephone networks recover from failures on the order of milliseconds [2], the Internet routing table convergence has been observed to take a much longer time [3]; it is not uncommon to have tens of minutes of downtime before recovery has taken place. Although empirical observations of the dynamics have been made in [4, 5], formal models of the process are not very common (there has been recent work on an analytical study of Route Flap Damping [6], and another study of BGP in congested networks [7]). There

*This research is supported by DARPA Contract N66001-99C-8614, DARPA Contract F30602-00-554, NSF EIA-0080119 and ANI-0085848

†Corresponding author

are other models of BGP that verify correctness, satisfiability etc. [8, 9, 10], but in this paper we explore the dynamical behavior of BGP. Our paper is a step in trying to understand the resilience of networks, and is inspired by recent studies [1] that have indicated vulnerabilities in the routing infrastructure and have conjectured that cascading failures can occur due to the interactions in BGP. We develop and analyze models in this paper that reflect the behavior of the protocols. Our analysis confirms the observations made in [1], that it is indeed possible to have cascading failures in systems like the BGP routing infrastructure. Our results indicate the presence of phase transitions in these systems, and the presence and intensity of the phase transitions are strongly dependent on system parameters.We use the term *phase transitions* in the sense used by Erdös and Renyi in their work on random graphs, that is "an abrupt change in a global system property". The phase transition is to be interpreted as a sharp threshold rather than the definition used in statistical mechanics. In our analysis, we observe that the propensity for phase transitions increases as clique size increases, and additionally also increases as the processing capacity of the routers decreases.

The rest of the paper is organized as follows: In Section 2 we give a brief description of the study in [1]. In the next Section, we develop a fluid model for the interactions and study its behavior. Next, in Section 4 we develop a birth-death model for the system, refining the efforts in the previous section. In Section 5 we analyze the behavior of the birth-death model and comment on the behavior of the system. Finally, we conclude in Section 6.

2 Background

An interesting vulnerability of BGP came to light in July and September of 2001. This was the incident where CODE RED and Nimda viruses disrupted the Internet routing infrastructure and caused widely reported BGP storms [1]. We briefly describe the incident, discuss possible explanations and then elucidate how modeling can help in this scenario. A direct quote from [1] is reproduced below

> *On July 19th, we observed an exponentially growing eight-fold increase in the advertisement rate, over a period of about eight hours (all times are in GMT; subtract 4 hours for EDT). This BGP surge faded over the same time scale as it arrived. When one considers the conventional wisdom about BGP convergence times (seconds to minutes), it is more than a little disturbing to see a fundamental quantity like BGP advertisement rate exhibiting exponential growth for eight hours.*

This event coincided with the spread of the CODE RED virus on the Internet. Then again, on September 18th, the day the NIMDA virus started to spread, the following was observed:

> *On Tuesday, September 18, simultaneous with the onset of the propagation phase of the Nimda worm, we observed another BGP storm. This one came on faster, rode the trend higher, and then, just as mysteriously, turned itself off, though much more slowly. Over a period of roughly two hours, starting at about 13:00 GMT (9am EDT), rrc00 [1] aggregate BGP announcement rates exponentially ramped up by a factor of 25, from 400 per minute to 10,000 per minute, with sustained "gusts" of more than 200,000 per minute. The advertisement rate then decayed gradually over many days, reaching pre-Nimda levels by September 24th.*

[1] rrc00 is a RIPE NCC site in Amsterdam, Holland

These two events clearly demonstrated that an application layer event (the virus attack) caused problems at lower layers of the Internet infrastructure. What caused this unexpected event? Briefly, [1] hypothesized the following chain of events

- The viruses started random IP port scanning
- Most of these random IP addresses were not in the cached entries of the routing table, causing....
- Frequent cache misses, and..
- In the case of invalid IP addresses, generation of ICMP (router error) messages..
- Both of the above causes led to router CPU overload, causing routers to crash
- Router failure led to withdrawal announcements by the peers, generating a high level of advertisement traffic.
- When the router came back on, it required a full state update from it's peers, creating a large spike in the load of it's peers who provided the state dump
- Once the restarted router obtained all the dumps, it dumped *its* full state to all its peers, creating another spike in the load..
- Frequent full state dumps led to more CPU overload, leading to more crashes, and the propagation of the cycle...

An immediate question of concern is: how self-sustaining is this process? On the two particular days, the BGP storms lasted for hours, causing major disruptions on the Internet. Although there is some debate raging over the findings of [1], e.g. [11], our interest is answering the question if this process (or something similar) can result in cascading failures. In the next two sections, we develop simple models to analytically understand this behavior.

Before we begin our modeling, we give a brief overview of BGP behavior. BGP is the inter-autonomous system (AS) routing protocol. At the boundary of each autonomous system, peer border routers exchange network reachability information with other autonomous systems through BGP. BGP uses TCP as its transport protocol. Two BGP speakers form a transport protocol connection between one another, and they exchange messages to open and confirm the connection parameters. When a connection is first established, a BGP speaker sends its entire routing table to the peer (a full state dump). During the following BGP session, incremental updates are sent as the routing table changes. Two types of BGP messages are important to BGP operation. First is KeepAlive message, which is sent periodically to ensure the connection is live. If the peers can't receive KeepAlive messages in a preset period of time, the BGP connection has to be closed. Physical connectivity failure (link failure, router crash), transient connectivity problems due to congestion, or even manual reboots, may result in the delay of KeepAlive message to the peers. When BGP sessions restart, the peers have to send the full routing table again. Update messages are used to exchange routing information change between two peers. Route withdrawals are sent when a router makes a new local decision that a network is no longer reachable. We study the clique topology (fully connected mesh) in subsequent sections. The clique topology provides the most complex interactions, and in the core of the Internet the major Autonomous Systems form a clique.

3 A Fluid Model

Let's build a very simple model of the interactions in the system and explore its properties. We consider a simple scenario with a finite set of N routers, that are all connected to each other, i.e. form a clique. Let the number of down routers at any instant be $D(t)$. We define a "down" router to be one that does not have a functioning routing table, so a router that is in the process of rebooting and obtaining state dumps is also defined to be "down". Now, we study the system of the number of down routers. The arrival and departure process to the system is defined as:

$$\alpha(t) \doteq \text{Number of arrivals in } [0, t) \tag{1}$$

$$\delta(t) \doteq \text{Number of departures in } [0, t) \tag{2}$$

Consider the process $\delta(t)$. The down routers come up with the help of the routers that are currently up $(N(t) - D(t))$. We define the service rate of an up router as k_s, where k_s is the average number of down routers a functioning router restores per unit time. Now, if $N(t)-D(t)$ servers are up and providing service, the service received by a single down node is its share of the total service capacity of the system. Thus, the share received by a single router is $(N(t) - D(t))/D(t)$. To account for the boundary condition $D = 0$, the denominator should actually be the the term $(D(t) + k_a)$, where k_a represents the ambient load on the servers, representing for instance processing of normal route advertisements, and prevents $D \to -\infty$. However, for simplicity of exposition we ignore the term as it does not affect the main observation we obtain later. Hence, the number of departures in an infinitesimal time dt, $d\delta(t)$, is defined by

$$d\delta(t) = D(t)\frac{N(t) - D(t)}{D(t)}k_s dt = (N(t) - D(t))k_s dt \tag{3}$$

Resetting of BGP sessions lead to two kinds of messages, withdrawal announcements and subsequently full state updates when the BGP session is restored. We model the rate at which a functioning router goes down due to the load imposed by the resetting of a single BGP session as k_l. Typically, we expect $k_s >> k_l$, as BGP resets are not uncommon and restoring a single session is unlikely to cause a peer router to go down. Now, the average arrivals in an infinitesimal time dt, $d\alpha(t)$, is given by the product of three quantities: the constant k_l, the number of routers (BGP sessions) that are currently down $D(t)$ (denoting the total load offered), and finally the number of routers currently up (that can go down), $(N(t) - D(t))$, i.e.

$$d\alpha(t) = k_l D(t)(N(t) - D(t))dt \tag{4}$$

Now $D(t)$ is $\alpha(t) - \delta(t)$, hence combining (3) and (4) we obtain the drift equation

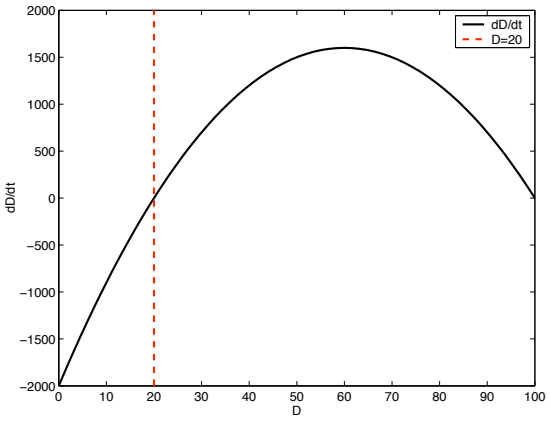
$$dD(t) = D(t)(N(t) - D(t))k_l dt - k_s D(t)\frac{N(t) - D(t)}{D(t)}dt \tag{5}$$

Simplifying and dividing by dt both sides, we obtain the following relation

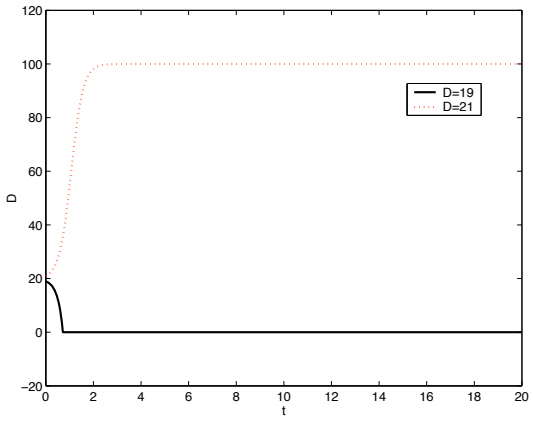
$$\frac{dD}{dt} = -k_l D^2 + (k_s + k_l N)D - k_s N \tag{6}$$

This is a Riccati equation, and without going into the actual solution of the equation, we immediately observe that the dynamical system described by this model exhibits a *phase transition*: If the initial state $D(0)$ of the system is above a certain threshold, then as $\lim_{t\to\infty} D(t) = N$, else $\lim_{t\to\infty} D(t) = 0$. In other words, if by some exogenous process (e.g. CODE RED)

we manage to bring a certain number of the routers down, thereby resetting the BGP sessions, then depending upon that number the system either fully recovers or there is a *cascading failure*. This is seen in Figure 1(a), where we plot the RHS of (6) as a function of D. This is very similar to results obtained in epidemic modeling, see [12]. For a value of D above k_s/k_l, the derivative of D remains uniformly positive, making the system drift to the state where all routers are down. For D below that number the derivative remains uniformly negative [2], bringing the system back to recovery. An interesting fact here is that this threshold is an *absolute* quantity rather than a fraction of N. Hence, given the right parameter set this phase transition may not be exhibited at all. A simulation of the system with different initial conditions is shown in Figure 1(b), where $k_s/k_l = 20$ and we plot two trajectories, one with $D(0) = 21$ and another with $D(0) = 19$. The presence of phase transition in the fluid model piques our



(a) Plot of dD/dt as a function of D



(b) Divergent behavior with different initial conditions

Figure 1: Phase transition in model

interest enough to attempt a refinement of the model, and observe the behavior. In the next section, we move away from the fluid model and develop a birth-death model for the system.

4 A Birth-Death model

Now we construct and analyze the discrete analog of the basic fluid model we studied in the previous section. Again, we assume there are N identical routers in the system, forming a clique. A state in the birth-death represents the number of down routers in the system. The model is depicted in Figure 2. The transition rate from state i to $i+1$ is given by $\lambda_i = (N-i) \times (i) \times k_l + k_a, i = 0 \ldots N-1$; and similarly the transition from $i+1$ to i is given by $\mu_i = (N-i) \times k_s, i = 0 \ldots N-2$. μ_{N-1} is defined to be zero, as there is no repair when all routers are down, hence state N is an *absorbing* state. To understand the behavior of this model, we have to perform a transient analysis, as the system ends up in state N with probability 1. The mean time to absorption from state i to N, W_i, is a good indicator of the behavior of the system. Physically, W_i is to be interpreted as roughly the time it takes for the system to collapse if i routers are currently down. Next, we compute W_i. In state i, the mean time to the next transition out of state i is $1/(\lambda_i + \mu_{i-1}), 0 < i < N$. At that transition, the

[2] If we account for k_a, the ambient load as earlier stated, then dD/dt is negative below the threshold except for $D = 0$, where it is 0. That does not change the qualitative nature of the result.

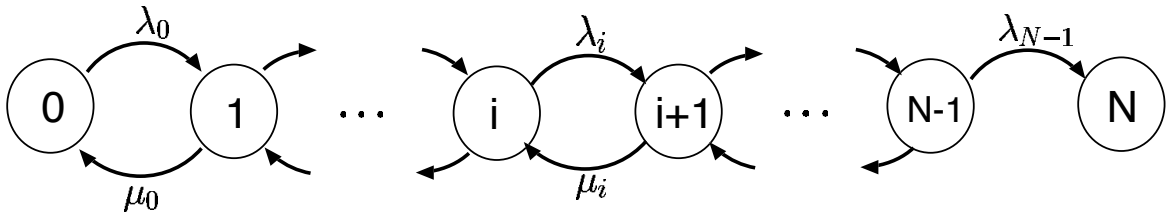


Figure 2: Birth Death Process: State i represents number of down nodes

expected remaining time to absorption is W_{i+1} with probability $\lambda_i/(\lambda_i + \mu_{i-1})$ and W_i with probability $\mu_{i-1}/(\lambda_i + \mu i - 1)$. Then

$$W_i = \frac{\mu_{i-1}}{\lambda_i + \mu_{i-1}} W_{i-1} + \frac{\lambda_{i-1}}{\lambda_i + \mu_{i-1}} W_{i+1} + \frac{1}{\lambda_i + \mu_{i-1}} \tag{7}$$

Next, we apply the boundary conditions that are $\mu_{-1} = 0$, and $\lambda_N = 0$. The first condition yields

$$W_0 = \frac{1}{\lambda_0} + W_1 \tag{8}$$

while the second boundary condition yields

$$W_{N-1} = \frac{1}{\lambda_{N-1} + \mu_{N-2}} + \frac{\mu_{N-2}}{\lambda_{N-1} + \mu_{N-2}} W_{N-2} \tag{9}$$

Combining equations (7), (8) and (9), we obtain for $i \geq 1$

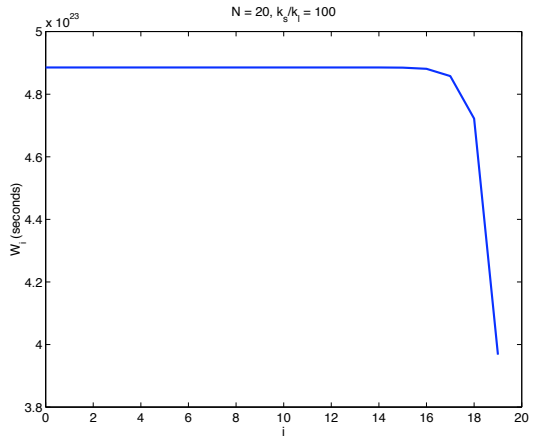
$$W_i - W_{i+1} = \sum_{j=0}^{i-1} \frac{\Pi_{k=j}^{i-1} \rho_k}{\lambda_i} + \frac{1}{\lambda_i} \tag{10}$$

Where $\rho_k = \lambda_k/\mu_k$. With $W_N = 0$, (10) yields a way to compute W_i as a function of i.

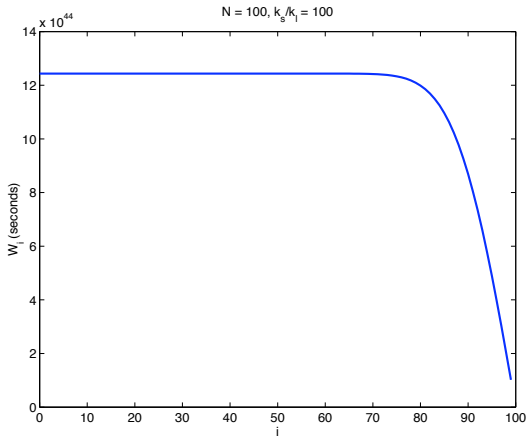
5 Model Analysis

The phase transition that we encountered in the fluid model, is manifested in the form of a sharp decrease in W_i in the birth-death model. We explore the parameter space of N, k_l, k_s and k_a to identify possibilities of cascading failures. First we assume that the "ambient" load, k_a, is a small fraction of the full state update load k_l. We choose the fraction to be $1/100$. Now we vary the ratio k_l/k_s and the size of the system N and observe the behavior. For k_l, we assume a value $0.01/s$, i.e., the rate of crashing of a router under the load of a full state update is 0.01 per second, and for k_s we assume a value of $1/s$, in other words a working router brings up a crashed router at the rate of 1 router per second. First we assume a small clique, of 20 routers. The mean time to absorption is shown in Figure 3(a). As we can observe the mean time to absorption is of the order 10^{23} seconds, as long as even one router is up. Hence, this system is unlikely to undergo cascading failures. Next, we increase the size of the clique, from 20 to 100, and plot the results in Figure 3(b).

The mean time to absorption now starts to show a decrease as i increases, but is still relatively quite high throughout. It is of the order 10^{44} throughout, hence when the system appears



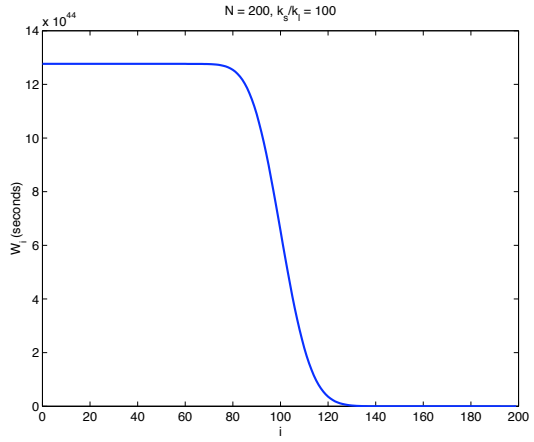
(a) Mean time to absorption for a small clique



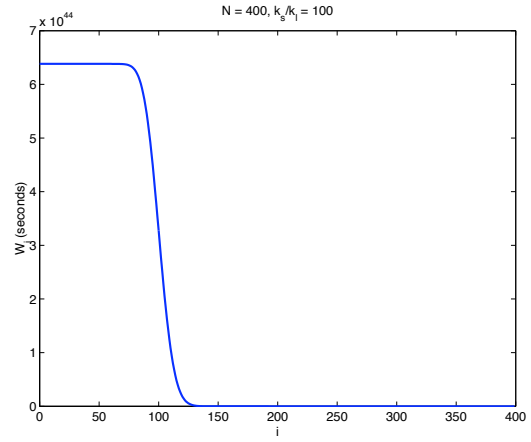
(b) Mean time to absorption for a medium clique

Figure 3: Stable behavior exhibited by small to medium cliques

to have become more stable, as the meantime to absorption has increased by several orders of magnitude. This makes intuitive sense, as increasing the size of the clique increases the redundancy in the system and there are more up routers "available" to bring up crashed routers. Now, we increase the clique size to 200, and show the results in Figure 4(a). Now, we observe the first appearance of the *phase transition*. At roughly $i = 100$, the mean time to absorption starts decreasing rapidly, and in a matter of few states, the time falls from an order of 10^{44} to close to 0! Thus, an apparently highly stable system becomes extremely vulnerable and can cascade into a collapsed state, where all routers are down in no time. This behavior is again observed in a clique of size 400, shown in Figure 4(b). The transition appears at around the same spot, around state 100.



(a) Mean time to absorption for a large clique (200 nodes)



(b) Mean time to absorption for a larger clique (400 nodes)

Figure 4: Phase transitions observed in larger cliques

The state 100 around which the phase transition starts to take effect is roughly the ratio k_s/k_l. Note that is similar to what we observed in the fluid model. This can be observed in Figure 5, where we plot the mean time to absorption for an identical clique size of 200, but vary the ratio k_s/k_l.

The effect of the clique size on cascading failures is also interesting. We now fix the ratio

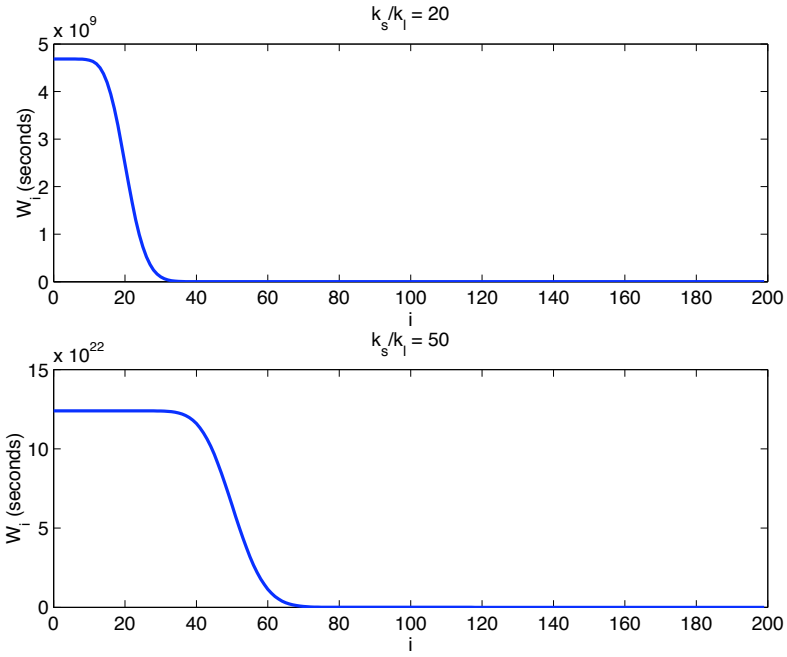


Figure 5: Phase transition point as a function of k_s/k_l

k_s/k_l, and vary N. The reults are plotted in Figure 6. Increasing N does not change the location of the phase transition, but does affect the relative stability before the transition. As we increase N, beyond the size where the system shows a phase transition, the relative stability *decreases*. This is in contrast to the scenario we earlier saw, where increasing N increased the relative stability for small to medium clique sizes.

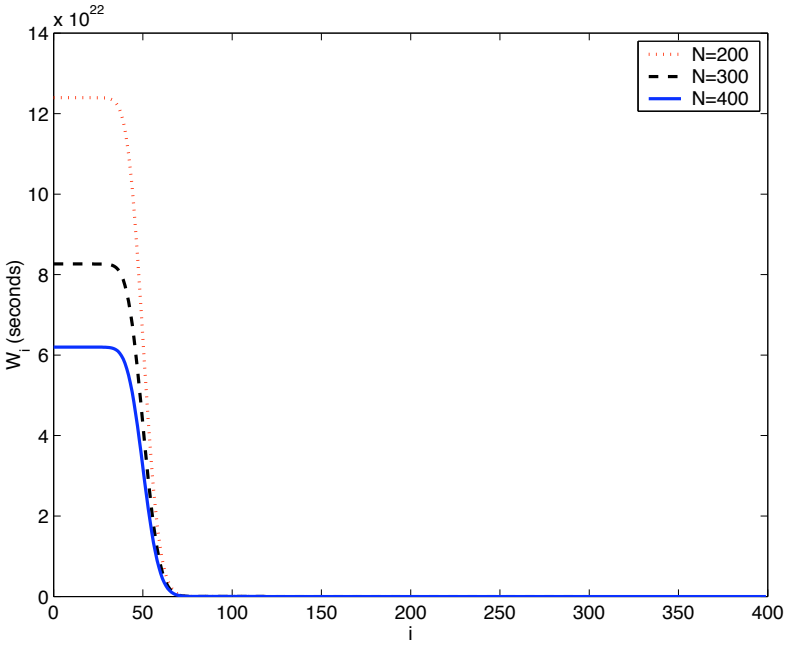


Figure 6: Relative stability and Phase transition point as a function of N

This behavior is more clearly observed in Figure 7, where we plot the mean time to absorp-

tion as N goes from 80 to 100 in steps of 10. We observe that the relative stability increases as N increases.

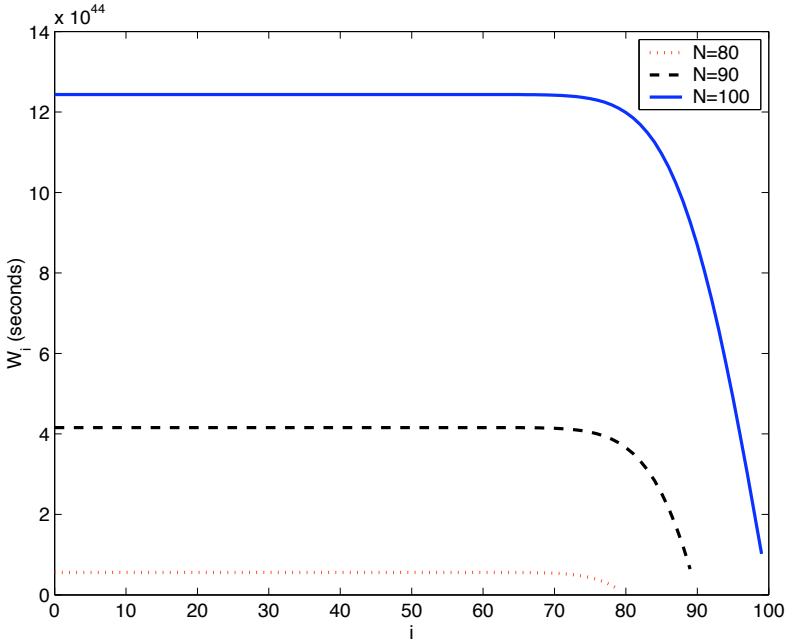


Figure 7: Relative stability and Phase transition point as a function of N

6 Conclusions

In this paper we have developed simple analytical models to capture the interaction of systems like BGP routing. We specifically develop models for the clique topology (the one that is used at the AS level on the global Internet) and discovered phase transitions with respect to cascading failures in the models. Our findings confirm the measurement based inferences made in [1] on possibilities of cascading failures in the routing infrastructure. Our models indicate that both the size of the clique as well as the capacity of the nodes in the clique is an important consideration for the phase transitions. The size of the clique acts as a threshold for the phase transitions, given other parameters, the clique must be large enough for the transition to appear. Increasing the clique sizes beyond the threshold does not change the location of the phase transition, but does have an effect on relative stability. On the other hand, if the clique size is large enough, then the capacity of the nodes in the system decides the location where the phase transition occurs.

In terms of future work, we are working on a more detailed model for a single router to understand individual node failures. A better understanding of individual node behavior would then permit us to model more complex topologies, moving beyond the homogeneous cliques that we have studied in this paper. An interesting question that our models can answer is to what kinds of interactions make the system most resilient? In other words, the transition rates in our model assume a certain kind of interaction. If that interaction is modified, does that lead to more resilient networks? That question of course has to be asked in conjunction with the issue of the performance of system under normal circumstances, and these are questions that we are trying to answer.

References

[1] J. Cowie, A. Ogielski, B. Premore, and Y. Yuan, "Global Routing Instabilities during Code Red II and Nimda Worm Propagation," http://www.renesys.com/projects/bgp_instability, September 2001.

[2] B. R. Hurley, C. J. R. Seidl, and W. F. Sewell, "A Survey of Dynamic Routing Methods for Circuit-Switched Traffic," *IEEE Communications Magazine*, vol. 25, no. 9, September 1991.

[3] K. Varadhan, R. Govindan, and D. Estrin, "Persistent route oscillations in inter-domain routing," *Computer Networks*, vol. 32, no. 1, pp. 1–16, Jan. 2000. [Online]. Available: http://www.elsevier.com/locate/comnet

[4] C. Labovitz, A. Ahuja, A. Abose, and F. Jahanian, "An experimental study of delayed internet routing convergence," Stockholm, Sweden, Aug. 2000. [Online]. Available: http://www.acm.org/sigcomm/sigcomm2000/conf/paper/sigcomm2000-5-2.pdf

[5] C. Labovitz, A. Ahuja, R. Wattenhofer, and S. Venkatachary, "The impact of internet policy and topology on delayed routing convergence," Anchorage, Alaska, Apr. 2001. [Online]. Available: http://www.ieee-infocom.org/2001/paper/507.ps

[6] Z. M. Mao, R. Govindan, G. Varghese, and R. Katz, "Route Flap Damping Exacerbates Internet Routing Convergence," in *Proceedings of ACM/SIGCOMM*, 2002.

[7] A. Shaikh, L. Kalampoukas, R. Dube, and A. Varma, "Routing stability in congested networks: Experimentation and analysis," in *Proceedings of ACM/SIGCOMM*, 2000. [Online]. Available: citeseer.nj.nec.com/shaikh00routing.html

[8] T. Griffin and G. T. Wilfong, "An analysis of BGP convergence properties," in *SIGCOMM*, 1999, pp. 277–288. [Online]. Available: citeseer.nj.nec.com/333036.html

[9] L. Gao and J. Rexford, "Stable internet routing without global coordination," in *Proceedings of ACM/SIGMETRICS*, 2000, pp. 307–317. [Online]. Available: citeseer.nj.nec.com/gao00stable.html

[10] L. Gao, T. Griffin, and J. Rexford, "On Inferring Automonous System Relationships in the Internet," in *Proceedings of IEEE/INFOCOM*, 2001.

[11] L. Wang, X. Zhao, D. Pei, R. Bush, D. Massey, A. Mankin, S. F. Wu, and L. Zhang, "Observation and Analysis of BGP Behavior under Stress," in *Internet Measurement Workshop*, 2002.

[12] D. Daley and J. Gani, *Epidemic Modeling*. Cambridge University Press, 1999.