Secure message passing on networks with insecure nodes

Sebastian M. Krause,[1] Michael M. Danziger,[2] and Vinko Zlatić[1]

[1] *Theoretical Physics Division, Rudjer Bošković Institute, Zagreb, Croatia*
[2] *Department of Physics, Bar Ilan University, Ramat Gan, Israel*

It is often necessary to transmit a message through a network when parts of the network are not secure. Here, we consider the case of a network partitioned into sets of nodes (colored with exactly one color) with the assumption that no single subset (color) can be trusted. One approach to obtain security is to divide the message and transmit the different parts on multiple paths so that no subset sees the entire message. This problem arises, for instance, in a peer-to-peer (p2p) network running different software versions. The same concern arises when considering the threat to anonymity posed by AS-level listeners in the entry and exit paths to the Tor network.

We present a general analysis of this problem based on a new kind of percolation theory and provide a framework to determine under what circumstances such an approach is feasible. In addition to analytic solutions for random networks and explicit verification for Erdős-Rényiand scale-free networks, we discuss a new kind of critical phenomena in which the critical exponent is determined by the number of subsets (colors) in the system.

Surprisingly, we find that increased software heterogeneity may actually improve security.

I. INTRODUCTION

Secure and anonymous communication over networks, in particular the internet, is a central question facing the global community. In light of widespread state-surveillance, cybercrime and severe bugs like "Heartbleed," one can no longer assume that an entire communications network is secure.

However, the network insecurity may be disjoint. For instance, on a p2p network, there may be different versions of the software running at the same time. In such a case, though there may be unpatched or even undiscovered bugs affecting a given version, it is unlikely that all of the versions will be compromised by the same group at the same time. In such a case, a sensible strategy for secure communication may be to divide the message and transmit it along different paths so that *no single version* receives the entire message. With this heuristic, secure communication may be achievable even if large parts of the network are insecure.

To analyze this problem, we consider a network for which each node is assigned exactly one color. We assume that one of the colors is insecure but *a priori* we do not know which one. The colors may stand for software versions on servers, where all servers of the same version are likely to fail at the same time; they may stand for ownership/control, where the controlling body (company, government etc.) may be assumed to eavesdrop on their nodes; they may stand for economic entities with correlated failure probability (due to financial dependence, reliance on the same resource); or they may stand for reloading points of transportation (e.g. ports with transferring goods from ship to train, where strikes could hit many ports at the same time).

The recent "Occupy Central" protests are a case in point. It was reported that mobile internet access was disrupted in an attempt to hamper the protesters' ability to organize. In response, many began to use a mobile application called "Firechat" which used Bluetooth to transmit messages to any other users within a few meters. Each user also automatically retransmits messages and in this way, given a sufficient density of users, local communicability can be recovered even in the absence of internet or phone service. However, it was also reported that government sponsored malware began to be spread via this network. Assuming not all operating systems or device manufacturers are equally susceptible, a new routing problem arises: can the messages be split so that no specific vulnerable class sees the whole message?

This problem also arises when attempting to safeguard anonymity with the Tor network [1]. Recent work has shown that if the same autonomous system (AS) controls a router on the path from the source to the entry node of the Tor network and also a router on the path from the exit node to the destination, the identity of source and destination can be deduced from a statistical analysis of the traffic pattern and the anonymity of the Tor network is broken [2]. Since a relatively small number of AS's control the entire internet, this scenario is a serious concern [3]. Indeed, we find that heterogeneity of management and versioning may provide higher levels of security if path heterogeneity is adopted.

a. Background How exactly do Dolev and Pinto relate to our problem? In the 1990s, using sets of paths with disjunct servers [4]. This early study, which gained broad attention in computer science [...], abstracted from the network structure and assumed the existence of the paths a priori. The possibility of secure communication was studied as well for wireless networks using percolation on spatially embedded graphs [5].

Here we examine the effect of network topology and number of subsets (colors) on the feasibility of secure communication on a network containing insecure sets of nodes. We begin with a formal definition of the problem and its relationship to percolation theory and then proceed to demonstrate a number of key properties on random graphs and sample measurements on real-world networks, including the AS-level internet.

On a non-colored network, if node or link failures occur with a given probability, percolation theory on complex networks can be used to determine overall connectivity [6, 7]. Here we develop a new framework based on percolation theory but not reducable to any previous percolation problems on complex networks.

II. AVOIDABLE COLORS PERCOLATION

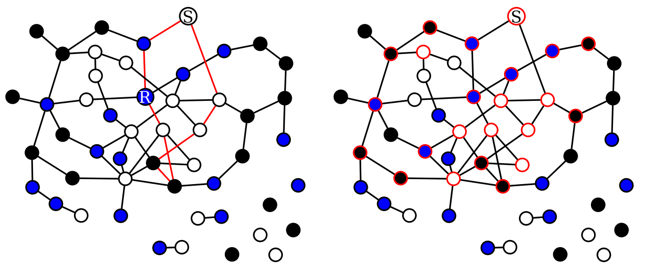

FIG. 1: Left: In this network the sender S and the receiver R can communicate with avoidable colors, as the short path highlighted with red avoids black and white nodes, and the long path avoids blue nodes. Right: All nodes highlighted with red belong to an avoidable colors component, as each pair out of this set is connected with avoidable colors. Notice that some nodes which are needed for connection of other nodes are not included in the component.

Assume a graph G with N vertices and adjacency matrix A_{ij}. Every vertex i has a color $c_i \in \{1, 2, \ldots, C\}$, where C denotes the total number of colors. Faced with the possible collective failure or insecurity of all nodes of a single color, "connectivity" means that two nodes are required to be connected while avoiding potentially vulnerable colors: For every color c, a connecting path must exist, such that *all* nodes on the path are *not* of color c. Since we must assume that the source and target themselves are secure, their colors are not included in the calculation of color-connectability. This is illustrated on the left of figure 1. In the following, we refer to this property as "color-connectability." We present a theory for the calculation of the probability of existence of a set of C paths, one avoiding each color, between pairs of nodes. There may be a smaller set of paths which securely connects the pair of nodes but such a case is trivially included in the condition of color-connectability via a maximal number of paths.

In order to discuss the connectivity of the network in general, we define a "color avoiding component" as a maximal set of nodes, where every node pair is color-connected. Such a component is highlighted with red in Fig. 1. Note that there are nodes which are not themselves part of the color avoiding component but they are necessary for the color-connected nodes to securely communicate. This occurs when all of the neighbors which lead from a node to the color avoiding component are of the same color. In such a case, the node itself is not color-connectable with the system as a whole because it must pass through nodes of a certain color before it can reach elsewhere. However, in general, this node will still be necessary to form paths which avoid other colors. The fact that non-color-connectible nodes may be needed to create overall system color-connectivity indicates that a new kind of percolation theory is needed here.

By studying the color avoiding component, we obtain a clear quantitative measure of the feasibility of security through multiple-path routing and information on where those paths should be routed. Furthermore, this gives us a way to measure the effect of changes in network topology, link density and number of colors.

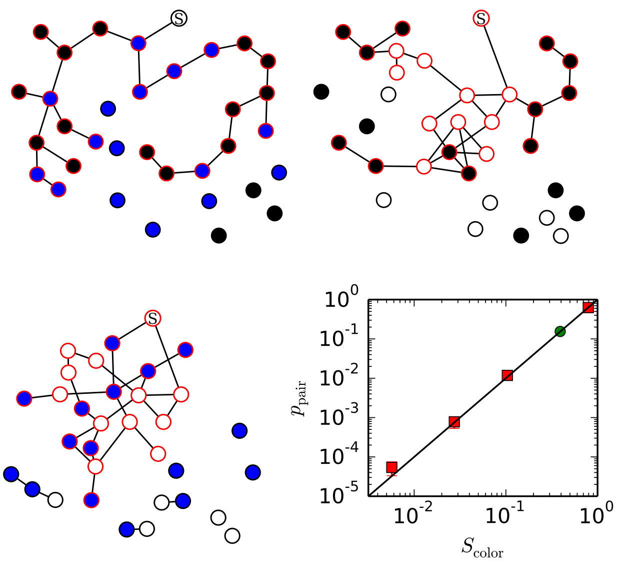

FIG. 2: Illustration of the construction of the set of nodes $\mathcal{L}_{\rm color}$ which is the largest avoidable colors component *I don't understand. Why many? in many large networks.* The largest components without white ($\mathcal{L}_{\bar{1}}$), without blue ($\mathcal{L}_{\bar{2}}$) and without black nodes ($\mathcal{L}_{\bar{3}}$) are highlighted in red, and the test node S is connected to all of them and therefore belongs to $\mathcal{L}_{\rm color}$. Lower right: Estimation of the fraction of successful pairs for quenched graphs with different values $S_{\rm color} = N_{\rm color}/N$. Red squares show Poisson graphs with increasing degree and $C = 3$ colors, the green circle shows the autonomous systems network with $C = 2$ colors. The black line indicates the case where only node pairs in $\mathcal{L}_{\rm color}$ are connected with avoidable colors. As numerical results are close, $\mathcal{L}_{\rm color}$ indeed dominates the secure communication abilities of many graphs. Notice that even for the smallest value shown, $N_{\rm color} = 570$ has reasonable size. The blue circle shows the network of autonomous systems with two colors distributed over the nodes. Our network snapshot of the year 2006 contains $N = 22963$ nodes. $p_{\rm pair}$ was approximated with samples of up to 5×10^5 pairs, error-bars are smaller than the symbols in most of the cases.

As illustrated in figure 2, there is a way to find a candidate set of nodes $\mathcal{L}_{\text{color}}$ for the largest avoidable colors component. First, for every color c, we delete all nodes with color c and find the largest component in the remaining graph, $\mathcal{L}_{\bar{c}}$. Next, we define $\mathcal{L}_{\text{color}}$ as the set of nodes such which are in $\mathcal{L}_{\bar{c}}$ or have at least one link to it; for every color c. Now every node pair in $\mathcal{L}_{\text{color}}$ are color-connected. *I don't understand why this makes it maximal...* If for every color c, $\mathcal{L}_{\text{color}}$ includes at least one node out of $\mathcal{L}_{\bar{c}}$, it is maximal and therefore it is an avoidable colors component. There is no easy way to test whether $\mathcal{L}_{\text{color}}$ is the largest avoidable colors component (as shown in figure 3, avoidable colors components can exist due to different mechanisms and they can largely overlap). However, we will see that $\mathcal{L}_{\text{color}}$ *This needs to be explained.* might scale with system size and in this case it can be considered as a giant avoidable colors component. Letting N_{color} equal the number of nodes in $\mathcal{L}_{\text{color}}$, we find that at least $N_{\text{color}}(N_{\text{color}}-1)/2$ out of all $N(N-1)/2$ possible node pairs in the network are connected with avoidable colors. This is a macroscopic fraction if $\mathcal{L}_{\text{color}}$ scales linearly with system size. We can use this fact to test whether $\mathcal{L}_{\text{color}}$ dominates the secure communication abilities of a network by plotting the fraction of pairs connected with avoidable colors in the whole network p_{pair} against $S_{\text{color}} = N_{\text{color}}/N$. In figure 2 on the lower right we see that secure connectivity is indeed dominated by $\mathcal{L}_{\text{color}}$. With red squares, results for Poisson graphs with $N = 10^5$ nodes and average degrees $\bar{k} = 1.6; 1.7; 1.9; 4.0$ are shown, where $C = 3$ colors were distributed over the nodes uniformly at random. *What are the other ways that color-secure communication can take place? Why does the deviation remain small? There's a piece missing here.* Results fit well even for small S_{color}. This validates the treatment of $\mathcal{L}_{\text{color}}$ as a proxy for color-connectivity and allows us to develop analytical results and understand the system's critical behavior, as discussed below.

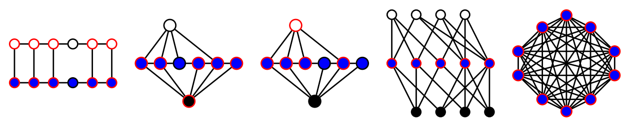

FIG. 3: These should probably be subfig-ed and labeled. Avoidable colors components, as highlighted with red, can be due to different scenarios. On the left, we see a *What makes this case scalable?* scalable case similar to random graphs. In the second graph, the high degree black node serves as an alternative paths provider for the blue nodes. In the third graph an alternative avoidable colors component is highlighted for that graph, showing that components might overlap. The second graph from the right does not need any connection among the blue nodes but there is a massive overhead of nodes and connections. On the right, we see that a clique is an avoidable colors component by definition.

To illustrate the rich phenomenology of avoidable colors components, some different mechanisms are shown in figure 3 which establish such components. On the left, we see a case which is similar to random graphs: all nodes which are neighbors to largest components without the color white and without the color blue can connect securely. In the second graph, the black node serves as an alternative paths provider for the blue nodes. It needs to have high degree for that. In the third graph an alternative avoidable colors component is highlighted. This shows that they might overlap and it is not straight forward to find the largest one. *Evidently, color-communication does not work with such components. How is this consistent with our understanding of the role of $\mathcal{L}_{\text{color}}$?* The second graph from the right does not need any connection among the blue nodes and the connecting white and black nodes have lower degree, however, there is a massive overhead of nodes and connections. On the right, we see a clique. In this case, no node of a different color is needed for all nodes to be color-connected, but the number of links needs to be maximal.

III. RESULTS

In this section we develop a new percolation theory to calculate the size of the maximal color-connected component. We present analytical results for the emergence and size of a finite-fraction S_{color} for random graph ensembles with randomly distributed colors, including critical phenomena, in the limit of infinite graphs. We confirm our analytical calculations with extensive numerical tests.

We begin with the generalized configuration model graph ensemble with N nodes, where each degree sequences $\{k_i\}$ occurs with probability $\prod_i p_{k_i}$ with the degree distribution p_k. Every node i is assigned a color $c_i \in 1, 2, \ldots, C$. For any degree sequence k_i, the color sequence $\{c_i\}$ has probability $\prod_i \tilde{r}_{c_i,k_i}$ with the degree-dependent color distribution $\tilde{r}_{c,k}$ ($\sum_c \tilde{r}_{c,k} = 1$ for every degree k separately). Do we really need to include a degree-dependent color distribution at this point?

We calculate S_{color} in the limit of $N \to \infty$ as the probability that a single node belongs to $\mathcal{L}_{\text{color}}$. This problem can be decomposed into two parts. First, all possible cases of neighborhoods are summed over with the according probabilities. Let κ_c be the expected number of neighbors of color c which are connected to the giant component of standard percolation. Calculating κ_c for all colors, we obtain the vector $\vec{\kappa} = (\kappa_1, \ldots, \kappa_C)$. Second, the conditional probability $P_{\vec{\kappa}}$ that these links suffice to connect to $\mathcal{L}_{\text{color}}$ can be calculated as follows:

$$P_{\vec{\kappa}} = \prod_{c=1}^{C} [1 - (U_{\bar{c}})^{\sum_{c' \neq c} \kappa_{c'}}], \tag{1}$$

$$P_{\vec{\kappa}} = \prod_{c=1}^{C} \left(1 - U_{\bar{c}}^{k' - \kappa_c}\right) \tag{2}$$

The second equation is from the supp. and includes k' - why does the main one not include k'? in which $U_{\bar{c}}$ (equation S...) denotes the conditional probability that a link fails to connect to $\mathcal{L}_{\bar{c}}$ given that it does connect to the normal

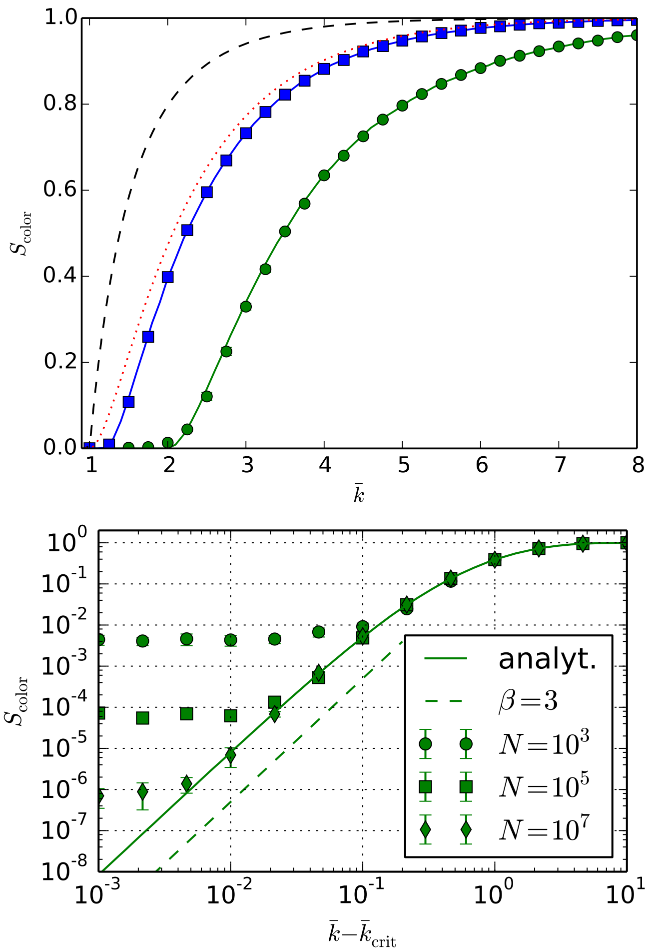

FIG. 4: Top: Dependence of $S_{\rm color}$ on average degree $\bar{k}$ for Poisson graphs with different numbers of colors. Symbols show numerical results for networks of size $N = 1000$ (blue squares for $C = 10$ and green circles for $C = 2$ colors), the straight lines show the corresponding analytical results. For comparison, the giant component size S is shown (black dashed). $S_{\rm color}$ is smaller than S via two mechanisms. First, every node has to be connected to the giant component via two links, as in 2-core percolation [8]. The corresponding fraction of nodes $S_{{\rm color},\infty}$ is shown with a red dotted line. Second, increasing color frequencies further decrease $S_{\rm color}$. Bottom: Finite size scaling for $C = 3$ colors emphasizes the dependence of the critical exponent β on the color distribution, here $\beta = C = 3$.

giant component via a node having a color $c' \neq c$. $U_{\bar{c}}$ is uniquely determined by [Eq 6 in Supplements?] We need to give more detail about $U_{\bar{c}}$ because it's not trivial or standard.

Using $P_{\vec{\kappa}}$, we can calculate the size of the largest-color connected component as

$$S_{\rm color} = \sum_{k=0}^{\infty} p_k \sum_{k'=0}^{k} B_{k,k'} \sum_{\kappa_1,\ldots,\kappa_C=0}^{k'} M_{k',\vec{\kappa}} P_{\vec{\kappa}}, \qquad (3)$$

where the binomial factor $B_{k,k'}$ (equation S...) accounts for the probability that out of k links k' links connect to the normal giant component. The multinomial factor $M_{k',\vec{\kappa}}$ (equation S...) gives the multinomial probability of having the color distribution $\vec{\kappa}$ among the neighbors belonging to the normal giant component.

Similar to standard percolation, we find that the size of the largest color-connected component $S_{\rm color}$ undergoes a phase transition at a specific $\bar{k}_{\rm crit}$. For $\bar{k} < \bar{k}_{\rm crit}$, color-connectability is confined to clusters of finite size (zero in the limit of large N) and for $\bar{k} > \bar{k}_{\rm crit}$ there is a maximal color-connected component $S_{\rm color}$ which scales with system size. We find that the value of $k_{\rm crit}$, decreases as C increases and approaches the standard percolation threshold as $C \to \infty$. The critical behavior is discussed in detail in the supplements, for the general case of heterogeneous color distributions. Approximations in equation 3 allow us to understand the critical behavior. Applied to the homogeneous color distributions discussed here, the results reduce to

$$S_{\rm color} \propto (\bar{k} - \bar{k}_{\rm crit})^{\beta} \qquad (4)$$

$$\beta = C, \quad \bar{k}_{\rm crit} = C/(C-1). \qquad (5)$$

We confirm the value of $\bar{k}_{\rm crit}$ and the scaling of $S_{\rm color}$ numerically in figure 4. For large numbers of colors, we observe the critical exponent β becomes large. *I don't understand this sentence. What about the size of the critical region? Isn't it significant that it converges to zero as the exponent diverges? How can we include that phenomenon concisely?* With this, the system shows an effectively shifted transition between vanishing and finite $S_{\rm color}$, as the growth of the giant avoidable colors component $\frac{\rm d}{{\rm d}k} S_{\rm color} \propto \beta(\bar{k} - \bar{k}_{\rm crit})^{\beta-1}$ is close to zero for small arguments. To our knowledge, this is a new kind of behavior, and a more detailed analysis of other quantities at the phase transition would be a fruitful topic for further research.

In Fig. 4 it is evident that, even as the number of colors tends to infinity, standard percolation *is not* recovered and $S_{\rm color}$ remains smaller than S. This is because nodes can only belong to $S_{\rm color}$, when they are connected to the normal giant component with at least two links. In other words, percolation in the limit of infinite colors is equivalent to k-core percolation with $k = 2$ [8]. To understand this we derive an asymptotic form for $S_{\rm color}$ as $C \to \infty$:

$$S_{{\rm color},\infty} = S - (1-u)\frac{dg_0(z)}{dz}|_{z=u} \qquad (6)$$

where $g_0(x)$ is the generating function of the graph and u is the probability that a link does not lead to the standard percolation giant component. Comparing this to Eqn (1) in Dorogovtsev et al. [8], we see that this is exactly the same equation as 2-core percolation.

In figure 4, we see that $S_{\rm color}$ comes close to $S_{{\rm color},\infty}$ even for $C = 10$. For the Poisson graph we show in the supplements that $S_{{\rm color},\infty} \propto (\bar{k}-1)^2$ which grows slowly for small parameter $\bar{k} - 1$.

$S_{{\rm color},\infty}$ represents the maximal size of the color-connected component. When there is a smaller number of colors, the finite color frequencies $\tilde{r}_{c,k}$ further reduce the percolating fraction of nodes. This also changes the critical value $\bar{k}_{\mathrm{crit}}$ and the critical exponent β. I think we need more discussion of the critical phenomena including some equations. Since this is supposed to be a major part of the paper, we need to support it better.

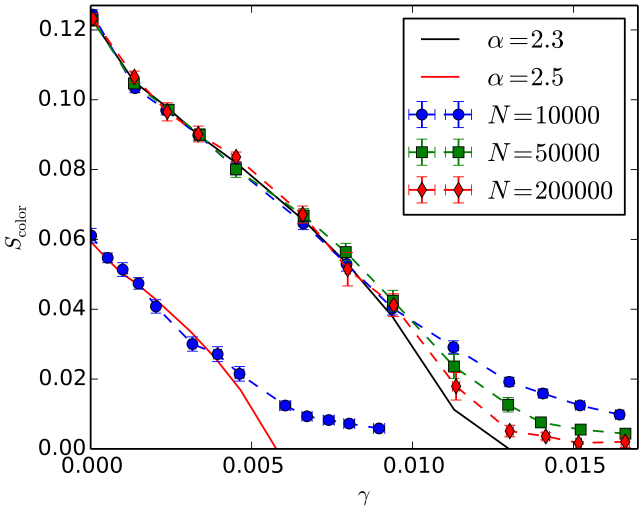

FIG. 5: On graphs with broad degree distribution, S_{color} drops fast, if a fraction γ of nodes with the highest degree is restricted to one color, while the nodes with smaller degree can have one of two colors. This is shown for networks with $\alpha = 2.3$ and $\alpha = 2.5$ having $N = 10000$ with symbols. Analytical results need the modified version of equation 3 as described in section **??**. Results are shown with straight lines. Without diversity on the hubs, nodes cannot communicate in the desired way.

Lets now discuss graphs with broad degree distributions with $p_k = nk^{-\alpha}$ with normalization constant n. For broad degree distributions, color distributions can show an additional type of heterogeneity, as a dependence of frequencies on the degree of a node can strongly influence the behavior. We used two colors, where the first color has a frequency of $\tilde{r}_{1,k} = 1$ for all degrees $k \geq k_{\mathrm{step}}$ larger than a certain k_{step}. These nodes have a probability of $\gamma = \sum_{k=k_{\mathrm{step}}}^{\infty} p_k$. Accordingly $\tilde{r}_{2,k} = 0$ for $k \geq k_{\mathrm{step}}$, and probabilities for smaller degrees are chosen as $\tilde{r}_{c,k} = 1/2$. Figure 5 shows results for an ensemble with $\alpha = 2.3$ and $\alpha = 2.5$. The analytical results for $\alpha = 2.3$ show that already for a portion of $\gamma = 1.4\%$ of the largest nodes occupied by the first color exclusively, S_{color} vanishes.

The red circles in figure 6 show results for the autonomous systems network, where colors where distributed with degree-dependence over the nodes as described at the end of the last section. Averages where taken over 10 realizations of the color distributions. As expected from our results for scale free degree distributions, S_{color} drops to 0 even for small fraction γ which is exclusively of one color. That means that if there is no heterogeneity in the highly connected servers, it is not

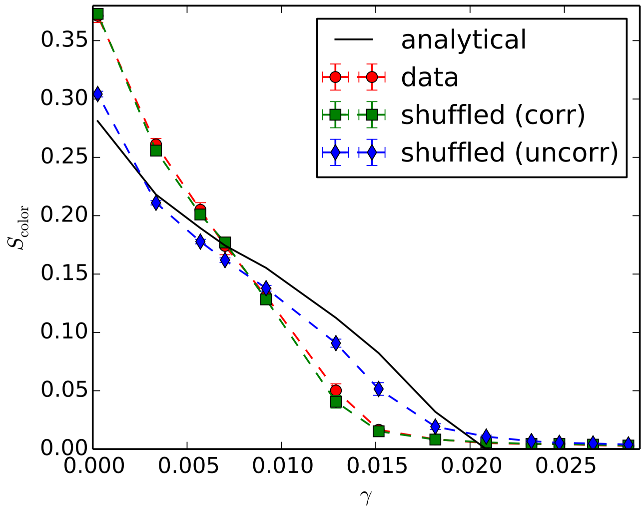

FIG. 6: Red circles show results for the network of autonomous systems, where colors are distributed in the same way as in figure 5. Averages were taken over 10 color distributions. Analytical results are shown with the black line and reproduce the results qualitatively. Deviations are due to degree-degree-correlations which are reserved in shuffled networks shown with green squares, while results with ignoring correlations are shown with blue diamonds.

possible to avoid e.g. software versions. This is also interesting in the following sense: It is known that secret services try to store all decrypted data running through servers to decrypt it later. As this is connected to technical afford, the services will more likely monitor the large servers. Therefore it would be beneficial, if once using encryption, to sent parts of the message using small servers. Unfortunately, this seems to be impossible, the services only have to monitor a low percentage of servers to hinder alternative paths.

In order to assess the predictive power of our analytical method, we used a model ensemble with using the degree frequencies of the autonomous systems network as degree distribution p_k. Results for the according ensemble are shown with the black line. The qualitative behavior is represented well. To understand deviations, we compared to data from shuffled networks starting with the original data. Shuffling with ignoring degree-degree-correlations while only keeping the degree sequence gives results close to the analytical results (blue diamonds). Shuffling with also keeping degree-degree-correlations gives results close to the original network (green squares). Therefore, deviations between our theory and the data arise mainly due to degree-degree-correlations.

IV. SUMMARY AND OUTLOOK

[1] R. Dingledine, N. Mathewson, and P. Syverson, in *Proceedings of the 13th Conference on USENIX Security Symposium - Volume 13*, SSYM'04 (USENIX Association, Berkeley, CA, USA, 2004) pp. 21–21.

[2] S. Murdoch and P. Zieliski, in *Privacy Enhancing Technologies*, Lecture Notes in Computer Science, Vol. 4776, edited by N. Borisov and P. Golle (Springer Berlin Heidelberg, 2007) pp. 167–183.

[3] M. Edman and P. Syverson, in *Proceedings of the 16th ACM Conference on Computer and Communications Security*, CCS '09 (ACM, New York, NY, USA, 2009) pp. 380–389.

[4] D. Dolev, C. Dwork, O. Waarts, and M. Yung, J. ACM **40**, 17 (1993).

[5] P. Pinto, J. Barros, and M. Win, Information Forensics and Security, IEEE Transactions on **7**, 125 (2012).

[6] R. Cohen and S. Havlin, *Complex Networks: Structure, Robustness and Function* (Cambridge University Press, 2010).

[7] M. Newman, *Networks: an introduction* (OUP Oxford, 2010).

[8] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Phys. Rev. Lett. **96**, 040601 (2006).

[9] S. Tauro, C. Palmer, G. Siganos, and M. Faloutsos, in *Global Telecommunications Conference, 2001. GLOBECOM '01. IEEE*, Vol. 3 (2001) pp. 1667–1671 vol.3.

[10] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, Proceedings of the National Academy of Sciences **104**, 11150 (2007).