

Secure message passing on networks with insecure nodes

Sebastian M. Krause and Vinko Zlatić

Theoretical Physics Division, Rudjer Bošković Institute, Zagreb, Croatia

Michael M. Danziger

Department of Physics, Bar Ilan University, Ramat Gan, Israel

It is often necessary to transmit a message across a network when parts of the network are not secure. Here, we consider the case of a network partitioned into sets of nodes with the assumption that no single subset can be trusted. As such, the message needs to be divided and transmitted 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 unpatched software versions and when considering AS-level listeners in the entry and exit to the Tor network.

We present a general analysis of this problem on random graphs including analytic solutions for Erdős-Rényi and scale-free networks and numerical simulations confirming our calculations and further numerical tests on real-world networks including the internet, partitioned by AS.

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

I. INTRODUCTION

Secure and anonymous communication over networks, in particular the internet, has become a central question facing the global community. In light of widespread state-surveillance, large-scale cybercrime and superbugs 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 can be achieved even if large parts of the network are insecure.

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.

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 this problem on general network topologies several partitioning rules. We do not consider here the implementation details of such a communication strategy but rather show under what circumstances it would be possible in principle. 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.

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. Therefore, a pair of nodes is securely connectable iff there exist a set of paths connecting the nodes such that no color appears on all of the paths. This property suggests an analytic approach similar to percolation theory on networks. If node or link failures occur with a given probability, percolation theory on complex networks can be used to determine overall connectivity [6, 7]. **Do we use k -core percolation? If not, why do we need to discuss it?** In particular, k -core percolation [8] has interesting implications due to the special k -core structure of the AS server network of the Internet [9, 10].

To calculate the probability of secure communication, we develop a new kind of percolation theory in which topological connectivity alone is not a sufficient condition. Rather, a pair of nodes is connectable if there exists a set of paths between them, each of which avoids one of the colors in the system. 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 connectability via a maximal number of paths.

II. AVOIDABLE COLORS PERCOLATION

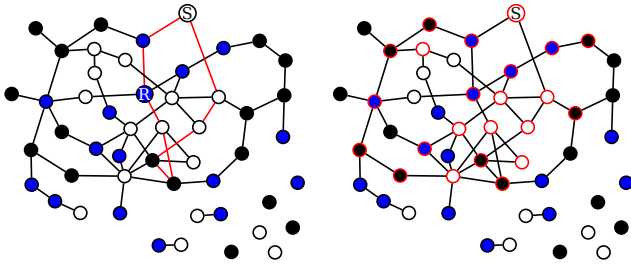


FIG. 1: Left: In this network the sender S and the receiver R can communicate with avoidable colors, as the short path highlighted with red avoids black and white nodes, and the long path avoids blue nodes. Right: All nodes highlighted with red belong to an avoidable colors component, as each pair out 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 $c_i \in \{1, 2, \dots, C\}$, where C denotes the total number of colors. 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). 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 with avoidable colors”: For every color c , a connecting path must exist, such that *all* nodes on the path are *not* of color c . This is illustrated on the left of figure 1. In the following, we refer to this property as “color-connected.”

In order to discuss the connectivity of the network in general, we define an “avoidable colors component” as a maximal set of nodes, where every node pair is connected with avoidable colors. Such a component is highlighted with red in Fig. 1.

Note that there are nodes needed for providing connections which themselves do not belong to the avoidable colors component. This is a very interesting property. We should discuss it more thoroughly.

By studying the avoidable colors 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, number of colors and the color distribution.

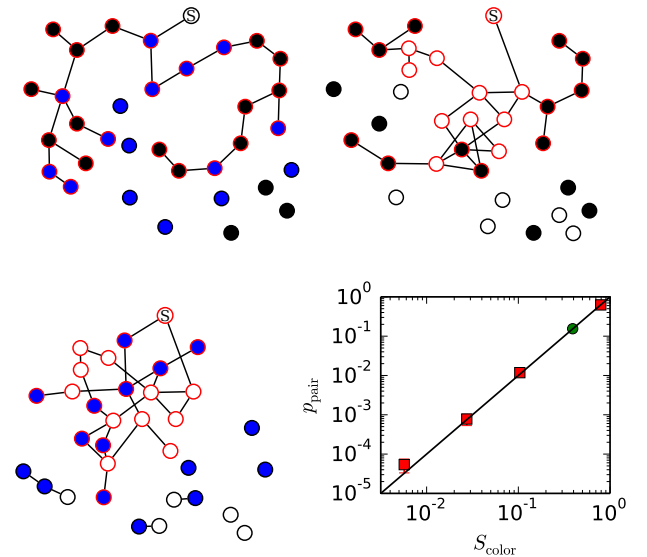


FIG. 2: Illustration of the construction of the set of nodes $\mathcal{L}_{\text{color}}$ which is the largest avoidable colors component *I don't understand. Why many? in many large networks.* The largest components without white (\mathcal{L}_1), without blue (\mathcal{L}_2) and without black nodes (\mathcal{L}_3) are highlighted in red, and the test node S is connected to all of them and therefore belongs to $\mathcal{L}_{\text{color}}$. Lower right: Estimation of the fraction of successful pairs for quenched graphs with different values $S_{\text{color}} = N_{\text{color}}/N$. 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 $\mathcal{L}_{\text{color}}$ are connected with avoidable colors. As numerical results are close, $\mathcal{L}_{\text{color}}$ indeed dominates the secure communication abilities of many graphs. Notice that even for the smallest value shown, $N_{\text{color}} = 570$ 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_{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.

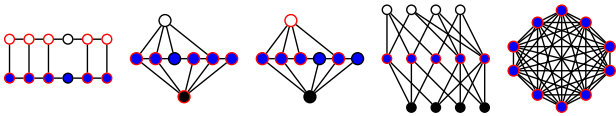


FIG. 3: *These should probably be subfig-ed and labeled.* Avoidable colors components, as highlighted with red, can be due to different scenarios. On the left, we see a *What makes this case scalable? scalable* case similar to random graphs. In the second graph, the high degree black node serves as an 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

We can find analytical results for random graph ensembles with randomly distributed colors in the limit of infinite graphs. The detailed derivation and explanation can be found in the supplements. We use 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 . Additionally we want to assign to every node i a color $c_i \in 1, 2, \dots, C$. For given 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).

We calculate S_{color} in the limit of $N \rightarrow \infty$ as the probability of a single node to belong 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. We need to specify the numbers κ_c for all colors c . κ_c counts the number of neighbors having color c and at the same time being in the normal giant component. Second, with the vector $\vec{\kappa} = (\kappa_1, \dots, \kappa_C)$, the conditional probability $P_{\vec{\kappa}}$ that these links suffice to connect to $\mathcal{L}_{\text{color}}$ can be calculated. This can be realized as follows:

$$S_{\text{color}} = \sum_{k=0}^{\infty} p_k \sum_{k'=0}^k B_{k, k'} \sum_{\kappa_1, \dots, \kappa_C=0}^{k'} M_{k', \vec{\kappa}} P_{\vec{\kappa}}, \quad (1)$$

$$P_{\vec{\kappa}} = \prod_{c=1}^C [1 - (U_{\bar{c}})^{\sum_{c' \neq c} \kappa_{c'}}]. \quad (2)$$

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. In the second equation, $U_{\bar{c}}$ (equation S...) gives the conditional probability that a link fails to connect to $\mathcal{L}_{\bar{c}}$, while this link fulfills the precondition of connecting to a node in the normal giant component having a color $c' \neq c$. Using $U_{\bar{c}}$, $P_{\vec{\kappa}}$ is the probability that for every color c at least one link succeeds to connect to $\mathcal{L}_{\bar{c}}$.

In figure 4 on the top, the dependence of S_{color} on the average degree is shown for Poisson graphs with different numbers of colors C ($\tilde{r}_{c, k} = 1/C$ is chosen homogeneous and independent of k). Comparing to the standard giant component size S (dashed black line in the figure), the percolation sets in at increasing \bar{k} with smaller numbers

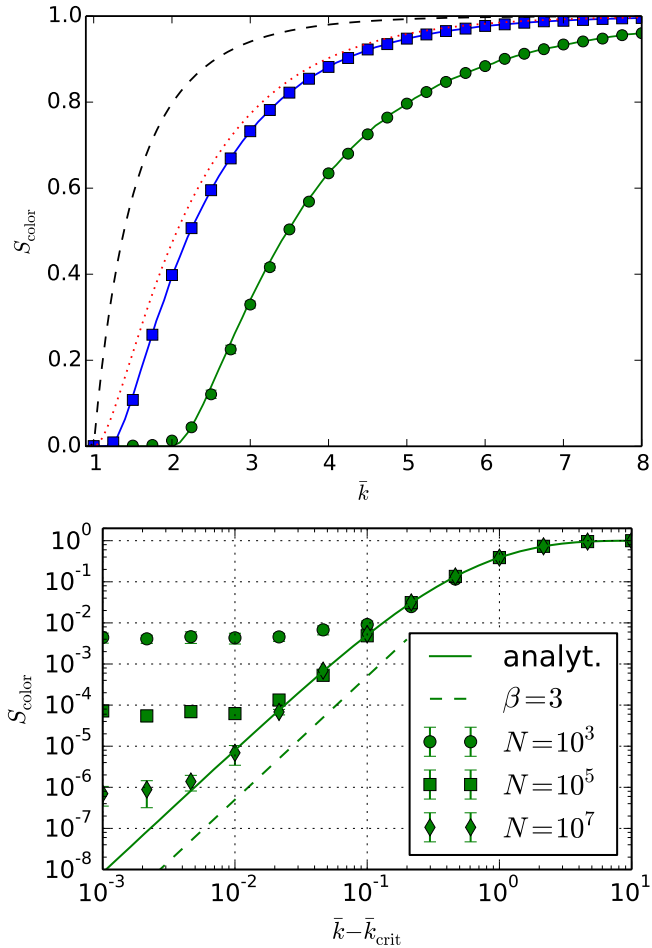


FIG. 4: Top: Dependence of S_{color} on the average degree for Poisson graphs with different numbers of colors. Symbols show numerical results for networks of size $N = 1000$ (blue squares for $C = 10$ and green circles for $C = 2$ colors), the straight lines show the according analytical results. For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

of colors, and the component size grows slower to the saturation value of one. The symbols show numerical results with $N = 1000$ and 100 network realizations, the lines show results of equation 1, both correspond well.

The suppression of the number of securely connected nodes can be understood as a combination of two effects. The first effect is purely topological and can be understood with $S_{\text{color},\infty}$ of eq. S?? (shown with dotted red line). It means that only nodes can belong to S_{color} , which are connected to the normal giant component over at least two links. We can confirm that S_{color} comes close

to $S_{\text{color},\infty}$ for high numbers of colors C with the results for $C = 10$. $S_{\text{color},\infty}$ is remarkably reduced compared to S for small k , but has the same critical parameter. For the Poisson graph we show in the supplements that $S_{\text{color},\infty} \propto (\bar{k} - 1)^2$ which grows slowly for small parameter $\bar{k} - 1$.

The second effect is connected to finite color frequencies $\tilde{r}_{c,k}$ which further reduces the percolating fraction of nodes. This also changes the critical value \bar{k}_{crit} and the critical exponent β . The critical behavior is discussed in detail in the supplements, for the general case of heterogeneous color distributions. Approximations in equation 1 allow us to understand the critical behavior. Applied to the homogeneous color distributions discussed here, the results reduce to

$$S_{\text{color}} \propto (\bar{k} - \bar{k}_{\text{crit}})^\beta \quad (3)$$

$$\beta = C, \quad \bar{k}_{\text{crit}} = C/(C - 1). \quad (4)$$

This behavior can be confirmed with numerical results. On the bottom of the figure, a finite size scaling for $C = 3$ colors establishes both the critical value of $3/2$ and the critical parameter. For large numbers of colors, we observe the interesting phenomenon of high critical exponent β . With this, the system shows an effectively shifted transition between vanishing and finite S_{color} , as the growth of the giant avoidable colors component $\frac{d}{d\bar{k}} S_{\text{color}} \propto \beta(\bar{k} - \bar{k}_{\text{crit}})^{\beta-1}$ 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 seems promising.

Lets now discuss graphs with broad degree distributions with $p_k = nk^{-\alpha}$. n is a normalization constant. 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 $k \geq k_{\text{step}}$ larger than a certain k_{step} . These nodes have a probability of $\gamma = \sum_{k=k_{\text{step}}}^{\infty} p_k$. Accordingly $\tilde{r}_{2,k} = 0$ for $k \geq k_{\text{step}}$, 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 were distributed with degree-dependence over the nodes as described at the end of the last section. Averages were 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 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

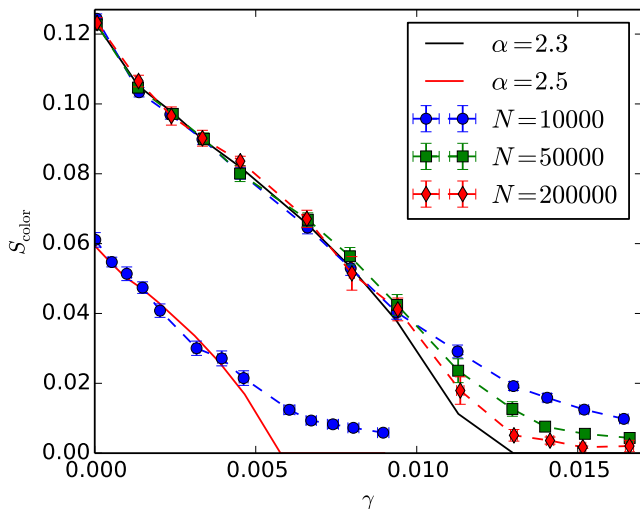


FIG. 5: On graphs with broad degree distribution, S_{color} drops fast, if a fraction γ of nodes with the highest degree is restricted to one color, while the nodes with smaller degree can have one of two colors. This is shown for networks with $\alpha = 2.3$ and $\alpha = 2.5$ having $N = 10000$ with symbols. Analytical results need the modified version of equation 1 as described in section ???. Results are shown with straight lines. Without diversity on the hubs, nodes cannot communicate in the desired way.

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. Dingleline, 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. Sigamos, 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).

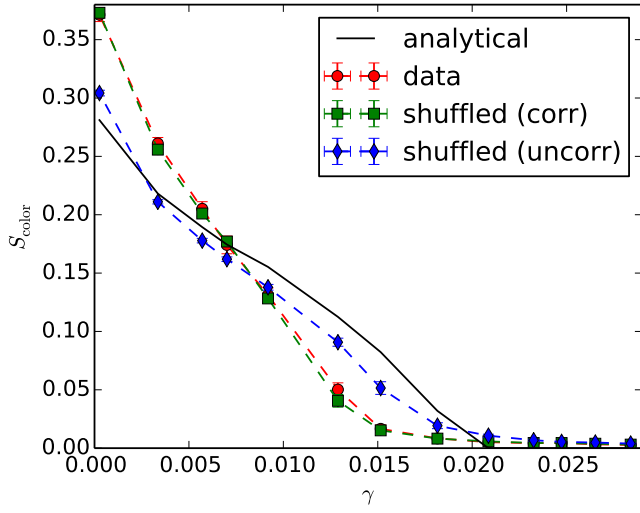


FIG. 6: Red circles show results for the network of autonomous systems, where colors are distributed in the same way as in figure 5. Averages were taken over 10 color distributions. Analytical results are shown with the black line and reproduce the results qualitatively. Deviations are due to degree-degree-correlations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