Providing Information Resilience through Modularity-based Caching in Perturbed Information-Centric Networks

Wei Koong Chai
Bournemouth University, Dorset, UK
wchai@bournemouth.ac.uk

Vasilis Sourlas, George Pavlou
University College London, London, UK
{v.sourlas, g.pavlou}@ucl.ac.uk

***Abstract*—In this paper, we investigate the provision of a new form of resilience, namely information resilience - targeting reliable communication of information under normal and adverse network conditions. We harness the power and flexibility of information-centric networking (ICN) paradigm where content are named and can be explicitly identified as opposed to the current host-centric Internet. Using ICN principles, a new modularity-based information caching approach is proposed that leverages the concept of modularity such that information reachability and persistency are enhanced especially under perturbed network scenarios, *e.g.*, network failures due to natural disasters or network under malicious attacks. The main idea of our proposal is to exploit the better connectivity of nodes within certain community construct in a network to provide higher information diversity, and thus allow potential access to a higher number of information objects even under various network dynamics. We conduct extensive simulations based on both real and model network topologies and show that our proposal can significantly increase request satisfaction ratios under highly dynamic network scenarios (*i.e.*, network under multiple perturbations) across different system parameters.**

I. INTRODUCTION

The society is increasingly expecting seamless information access at any moment and place. This expectation rides upon a plethora of recent technological advancements, culminating to the impending arrival of 5G networks, which promise ever-higher data rates (*i.e.*, tens of Gb/s peak rate) and shorter latencies (*i.e.*, sub-milisecond). In the heart of these is the key requirement to guarantee rapid and reliable response to information[1] requests regardless of device capabilities, bandwidth required and network conditions. While modern networks focus on low-latency communication, this is only one part of the total information response time. To ensure guaranteed response time, the design space should not be constrained by technological advancements in terms of speed only, but should also encompass real-time management of information (especially under network perturbations such as failures due to natural disasters, malicious attacks, *etc.*) for maximizing information resilience[2], which is the focus of this paper.

Even though the main aim of today's communication networks is to provide access to information, the current inter-networking paradigm still focuses on connecting physical host machines (*e.g.*, servers, fixed or mobile user devices, *etc.*), oblivious to the actual content being transmitted. Without understanding what is being communicated, there is no efficient way to discover alternative copies or information sources and thus, results in fragmentation of the Internet information space. In view of this, in the last few years the idea of information-centricity has emerged, where an information plane is formed in the network by naming information and treating information objects directly. This has been consolidated to the concept of information-centric networking (ICN) [1].

ICN directly addresses discovery, movement, delivery and management of information within a network while the underlying communication infrastructure is abstracted to allow flexible and seamless information access. ICN holds many attractive premises such as scalable, efficient, secure and flexible communication of information [2]. Most recently, it has been raised that information protection should be a first-order consideration within the network infrastructure [3]. By explicitly naming information, networks gain awareness of the information transported. ICN also changes the flow of information since the current end-to-end communication principle no longer holds. With its location-independence feature as well as in-network caching where information can be opportunistically cached in, identified and retrieved from any network node [4], information delivery is no longer restricted to the two communicating end-hosts.

We envisage a new form of resilience which focuses directly on the information rather than the network infrastructure. With information-centricity, we have the opportunity to provide resilience through its inherent anycast capability, whereby an information object can be delivered by multiple publishers or even caches. Failure of one information provider may not affect information delivery since the information is not strictly bounded to that specific host. Traditional network resilience focuses mostly on infrastructure resilience; protecting the underlying network infrastructure along with techniques such as load balancing and server duplication that protect servers. The newer delay-tolerant networking (DTN) approach offers connectivity resilience through its store-and-forward capability, combating the problems of intermittent connections.

This work focuses on studying *information resilience* whereby the focal point is on getting the requested information, independent of its location, to the user, regardless of the network and connectivity conditions. We build on the central information-centricity premises of ICN (*i.e.*, (1) network no longer agnostic to the information communicated, (2) natural

[1] The term *information* is used generically here referring to any data object produced, collected, shared or disseminated in the network. Furthermore, we use "information" and "content" interchangeably.

[2] We are interested in the resilience in communicating information in a network rather than resilience of information against intrusion or hacking; the latter relates to information security and is a different and important thematic research area in its own right.

anycast capability and (3) the ability of the network to cache and possibly manipulate information within network elements) and alleviate ourselves from relying on any specific ICN realization. We are interested in these common principles so that the results are insensitive to any ICN architecture and can even be applied to an information plane formed over the current host-centric Internet - for instance, through an overlay approach to create information awareness within that plane.

In this paper, a novel caching framework with regards to information resilience is proposed. The proposal exploits community structures within a network as the basic concept to enable information caching for resilience purposes. To the best of our knowledge, this work represents the first attempt to exploit community structures for information resilience which itself is a recently highlighted open issue in ICN. Prior work on in-network caching (See [5]) mostly focused on enhancing performance (such as improving cache hit ratio) while this work focuses on improving information availability and reachability taking into account possible failures in the network. For this, a review of the fundamentals of community structure as well as related work on resilience can be found in Section II. Our proposal is based on the concept of modularity and in Section III, we detail our solution and its possible instantiation/implementation. The evaluation of its performance across different impact factors can be found in Section IV, while a conclusion of the work can be found in Section V.

II. Background and Basics

A. Resilience and Caching in ICN

The topic on providing information resilience in communication networks is relatively new. However, there is a dense literature on network resilience[3], resulting in plenty of definitions and approaches. There is no commonly agreed terminology, *e.g.*, terms such as “robustness”, “resilience”, “dependability”, “survivability”, “reliability”, *etc.*, have all been used but often without clear differentiation amongst them. In [6], the authors created a systematic approach to the problem. Others, *e.g.*, [7], took a probabilistic view, arguing that since the behaviour of topological metrics, such as connectivity, is dependent on the characteristics of the network, then its robustness must also depend on these characteristics. Nevertheless, the work under this theme focuses mostly on protecting the physical network infrastructure (*e.g.*, maintenance of connectivity, fault detection, resource redundancy, *etc.*), such that the network can retain an acceptable level of functionality.

The explosive growth in information demands combined with evolving information dissemination patterns have highlighted the need for information identification to meet the requirements of future networks. With ICN, new solutions based directly on information can be developed to solve traditional IP networking issues. Along this line, [8] proposed a source recovery solution that utilizes knowledge of information to identify and thereby, recover the delivery process through alternative information sources when the original one fails. This goes beyond the scope of path recovery. With information named and identifiable, ICN opens the possibility of providing in-network information caching. We saw many recent proposals attempting to exploit in-network caching (see [5]) but they are almost universally aiming at metrics related to speed and efficiency of information delivery. Most recently, understanding the importance of reliable information transmission, [3] proposed a resilience scheme that exploits ICN in-network caching for content retrieval when the network is fragmented.

B. Community Structure

In this section, we briefly review the relevant basics of community detection in networks which forms the foundation of our proposal here. In general, there are two main components in community detections:

1) *Definition of community* – this questions what constitute a community and usually refers to the metric or criterion to determine community structure.
2) *Community formation algorithm* – this refers to the mechanics to find the community given a network.

While most initial work focused on graph cut size, here it is exploited the concept of *modularity* [9], [10] which is currently, by far, the most used function in determining communities in a network. Its definition has the advantage of embedding important components in community detection including community definition, choice of a null model as well as a measure on how strong the communities are [11]. The intuitive idea of modularity is that a community should have denser links within it compared to random link formation.

Let V, L and N denote the number of nodes, links and communities (or modules) in the network respectively. In [9], modularity of a network, Q, is measured as follows,

$$Q = \sum_i (e_{ii} - a_i^2), \qquad (1)$$

where $a_i = \sum_j e_{ij}$. Further, e_{ij} is the element in a $V \times V$ matrix E, representing the fraction of all edges in the network that link nodes in community i to nodes in community j.

Finding the maximum of the modularity metric has been shown to be an NP-complete optimization problem [12]. In this work, we explore two modularity algorithms. The first was described in [9] which from here onwards we refer it as *Newman_Girvan* algorithm. It is a lightweight technique to compute modularity, sacrificing some level of optimality. This method allows us to specify the number of resultant communities we would like and thus provides us with a “tuning knob”. Without going into details, *Newman_Girvan* finds communities by iteratively removing the link with the highest betweenness centrality with re-computation of link betweenness after each iteration.

The second algorithm we explored is proposed in [10] which from here onwards we refer it as *Newman_Eig*. It gives a good tradeoff between accuracy and scalability. In this algorithm, the principal eigenvector of the modularity matrix, B, is used. Given an adjacency matrix representing the network topology

[3] We use network resilience as a generic term that subsumes a loose and broad definition of any forms of resilience in communication networks against any kind of perturbations.

whereby its element A_{ij} is equivalent to 1 if there exists a link between node i and j and 0 otherwise, the modularity matrix can be written as follows

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2L}, \tag{2}$$

where $k_i = \sum_j A_{ij}$ is the degree of node i. The *Newman_Eig* algorithm iteratively splits the network based on the signs of the eigenvectors corresponding to the principal eigenvalue and terminates when all modules have largest eigenvalue equals to zero. The *Newman_Eig* algorithm detects *natural* community structures, arguing that some networks may not have a community structure at all and thus, forcing the algorithm to create specific number of communities may not be appropriate (*i.e.*, the *Newman_Girvan* algorithm).

Besides benchmarking our approach with the non-community-aware caching approaches (*e.g.*, the leave copy everywhere approach of NDN [13]; see Section IV), we also include an additional traditional graph partitioning algorithm, namely the spectral clustering algorithm [14] (shown as *Spectral* in our results) to illustrate the effectiveness of modularity-based algorithms (*i.e.*, *Newman_Girvan* and *Newman_Eig*) in our proposed solution compared to traditional approaches based on graph cut size. The *Spectral* algorithm is based around the observation that partitions of a graph with very low cut size can be obtained by assigning nodes based on the sign of the Fiedler vector (*i.e.*, the eigenvector corresponding to the second smallest eigenvalue of the Laplacian matrix).

III. Providing Information Resilience in ICN

A. Design Rationale

We argue that since communication networks are nowadays largely used for disseminating information, the information should be protected directly rather than focusing on the robustness of the physical network in the hope that the information will eventually be protected. In general, before any network failure happens, the network should be "prepared" so that access to different information objects will be least affected if and when the network is disrupted and after some network perturbations, we would like to protect any information objects that have their primary sources cut off from the network.

- **Before perturbations** – We seek to ensure maximal diversity of accessible information within certain topology constructs such that information availability and reachability are maximized. The idea is to exploit community concepts, *i.e.*, modularity here, since probabilistically, network fragmentations will most likely break a network between modules rather than within a module.
- **After perturbations** – We seek to protect any information that has lost its source from failures or network fragmentations such that information persistency is maximized. In this context, we would like to extend the availability/reachability of information (especially important for information such as announcements/news for public safety when natural disasters struck) such that such information stays reachable for as long as possible.

ICN enables information object to be identified individually based on names. This further allows in-network caching whereby network elements in the network are equipped with certain capacity to temporarily store information objects that traverse across them. A user's content request can then be served by any of the network element having a copy of the requested content. Such information retrieval mechanism follows a publish-subscribe approach though different implementations are found in different ICN architectures (*e.g.*, `Register` and `Find` primitives are used in [15], `Register` and `Interest` in [13] and `Publish` and `Consume` in [16], [17]). Here, we take such on-path caching mechanism as the starting point of our proposal. It is also assumed that the least recently used (LRU) cache eviction policy is used in all nodes.

B. Modularity-based Caching

The network operator, having the full knowledge of the network topology, computes the communities within its network. In this paper, we focus on modularity-based approach, specifically the *Newman_Girvan* and *Newman_Eig* algorithms. Each node will belong to one and only one community (*i.e.*, no overlapping communities). The network operator assigns a module ID to each community and configures each node within a community with the same ID (*i.e.*, all nodes within the same community are given the same ID).

For simplicity, for the rest of the paper, the primitives proposed in [13] are adopted using `Interest` as the content request, `Data` as the returned content and the description is based on the CCN/NDN router architecture though adapting the proposed approach to other ICN architectures is also straightforward. Specifically, the functionalities of three components from [13] are retained; namely the Content Store (CS) for storing cached content, the Pending Interest Table (PIT) for recording the `Interest` forwarded upstream towards the content source(s) and Forwarding Information Base (FIB) for forwarding `Interest` packets toward potential source(s) of matching `Data`. Briefly, a user requests a content by issuing an `Interest` packet which is forwarded based on FIB entries at each router, leaving a trail (*i.e.*, in PIT table). The `Data` packet is returned following the reverse of the path taken by the `Interest` packet. The pseudocode of the modularity-based caching algorithm in shown in Fig. 1, whereas in Fig. 2 and Fig. 3 are depicted the processing diagrams of an incoming `Interest` and `Data` packet respectively at an adapted intermediate NDN router.

The original NDN content router design is augmented with a new module ID table. This table will record the module ID attached in the `Data` packet that is received by the router. When the router decides to cache the packet, it stores the `Data` packet into its CS and correspondingly add a new entry in the module ID table to record its module ID. In the `Data` packet, a new module ID field is introduced to indicate from which module the content is originated from. This field is set by the first router that received the `Data` packet (see Line 17-18 of Fig. 1 and the top rhombus box in Fig. 3). Each router makes its own caching decision based on its own module ID and `Data`'s module ID. If the two IDs are different, it means that the information originated from outside the router's community. In this case, the router should cache

```
Interest – the received content request message
Data – requested content or information object
router(Mod_ID) – router's module ID assigned by the network operator
Data(Mod_ID) – module ID attached to the Data packet
IDF – Interest Destination Flag
SFC – Scoped Flooding hop counter
 1: if (Interest received) then
 2:     if Data in CS then
 3:         Return Data
 4:     else
 5:         if IDF==TRUE then
 6:             SFC += 1;
 7:             if SFC == max_SF_hop then
 8:                 Drop(Interest)
 9:             else
10:                 Forward(Interest, next_hop);
11:             end if
12:         else if IDF==FALSE then
13:             Forward(Interest, next_hop);
14:         end if
15:     end if
16: else if (Data received) then
17:     if Data(Mod_ID) == Null then
18:         Data(Mod_ID)=router(Mod_ID);
19:     else if Data(Mod_ID) != router(Mod_ID) then
20:         Insert(Data to CS);
21:         Insert(Data(Mod_ID) to Mod_ID_Table)
22:     end if
23:     Forward(Data, next_hop)
24: end if
```

Fig. 1: Pseudocode for the caching node

the content (see Line 19-22 of Fig. 1). Otherwise, if the two IDs matches, then based on our rationale that routers within the same community have significantly lower probability of being fragmented from each other, then the Data packet is not cached and the router simply forwards it to the next hop according to the matching PIT entries. This allows more cache space to store content from outside the community and thus increasing content diversity within each community. It is already know from previous works that limiting redundancy in caches are beneficial to the network as it improves metrics such as cache hit ratio [4].

As the focus here is on information resilience, we consider network perturbations whereby users may fail to get the requested content in the first attempt (*i.e.*, the initial Interest packet does not find the content source and no routers on its path has cached a copy of the requested content). In this case, the conventional approach of using a scoped flooding mechanism is followed whereby a user or a router sends a request which will then be broadcasted to all immediate neighbor routers. The broadcast of Interest should be limited to a small number of hops away from its origin to avoid high overhead. In addition, as studied in [18], if the content cannot be found within three hops away by scoped flooding, then the likelihood of finding the requested content further out is extremely low.

For this, an Interest Destination flag (IDF) bit is introduced to the Interest packet in order to distinguish whether the packet is heading towards the content origin (IDF is set to FALSE), following a FIB entry, or is heading towards neighboring routers following the scoped flooding resilience approach (IDF is set to TRUE). In the second case, a Scoped-Flooding Counter (SFC) is also introduced to keep track of the radius that such an Interest is transmitted. Refer to Line 5-14 of Fig. 1 and the diagram of Fig. 2 for the use of both IDF and SFC. In this paper, it is assumes that the FIB entries for all objects will be removed simultaneously from a router upon the disappearance/failure of a node along a path. This means that an Interest packet will have to visit only one network router (the one that the user is attached to) until the IDF is set to TRUE, upon fragmentation of the network. In case there is a delay for the update of the FIB entries, the proposed resilience mechanism can be enabled after a NACK is responded by the router which first identifies the content origin is unreachable or after the expiration of a timeout interval from the moment the user issued an Interest. The procedure followed by the network operator in order to detect a network fragmentation is

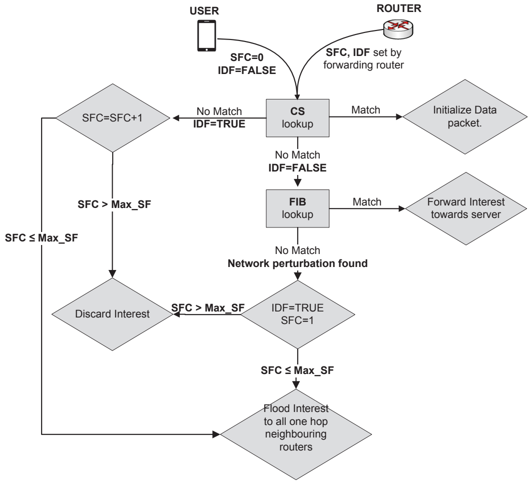
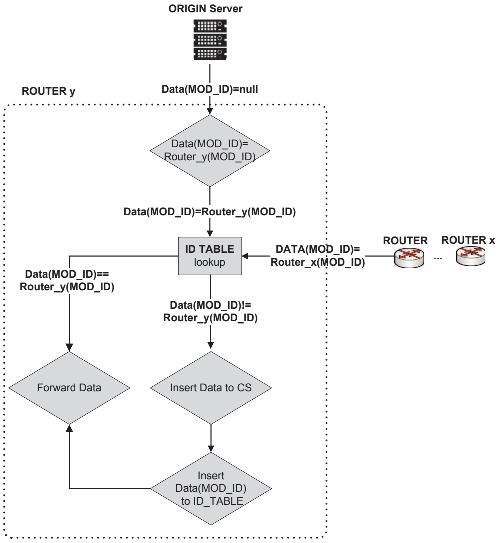


Fig. 2: Interest packet processing by an intermediate NDN router.



ORIGIN Server
ROUTER y
Data(MOD_ID)=null
Data(MOD_ID)=Router_y(MOD_ID)
Data(MOD_ID)=Router_y(MOD_ID)
ID TABLE lookup
DATA(MOD_ID)=Router_x(MOD_ID)
ROUTER
ROUTER x
...
Data(MOD_ID)==Router_y(MOD_ID)
Data(MOD_ID)!=Router_y(MOD_ID)
Forward Data
Insert Data to CS
Insert Data(MOD_ID) to ID_TABLE

Fig. 3: Data packet processing by an intermediate NDN router.

TABLE I: Summary of the topology and module properties (by *Newman_Eig*).

Network Properties	ER	SF	AT&T
Number of nodes V	100	100	113
Network Diameter	4	5	6
Link Density	0.0996	0.0594	0.0232
Number of Modules N	16	12	12
Mean Module Size	6.25	8.33	9.42
Median of Module Size	6.5	7.0	9.5
Maximum Module Size	11	22	18
Minimum Module Size	2	5	2
Range of Module Size	9	17	16

out of the scope of this paper. However, a scheme similar to the one presented in [8] for the PURSUIT architecture or the Named-data Link State Routing protocol (NLSR) presented in [19] for the NDN architecture could be adopted.

IV. Performance Evaluation

A. Simulation setup

We implemented the proposed modularity-based caching algorithm in a discrete event simulator. The simulator relies on a set of parameters which can be tuned to control the behaviour of the system. We experimented with both synthetic and real network topologies. For synthetic topologies, models based on the Erdős and A. Rényi (ER) random model [20] are used which has binomial degree distribution and scale-free (SF) graphs [21], which have power-law degree distribution. In the ER model, for a given number of nodes, a link randomly connects a pair of nodes with probability p_r independent of all other links. We constructed the ER graph with $V = 100$ nodes and $p_r = 2p_c = 2 \times ln(V)/V$ which is two times the sharp threshold for connectedness to ensure connected graph, while at the same time sufficiently small link density (link density = $L/\binom{V}{2}$) to avoid a highly meshed topology. For SF graph, the procedure described in [21] is followed with 3 links added based on the preferential attachment approach at each step. We also evaluated our proposal in the AT&T (AS 7018) network topology as provided through the Rocketfuel dataset (*i.e.*, 113 PoP routers) [22]. The properties of the topologies and the resultant modules are presented in Table I, while their degree distribution is shown in Fig. 4.

Each node is equipped with a cache of same capacity C, defined as a percentage of the entire content catalogue size, M. Request generation follows a Poisson process and its corresponding exponential distribution parameter is equal to 10 (*i.e.*, 10 requests/second from all the users in the network), whereas content popularity follows a Zipf distribution of slope z. Finally, it is assumed that each object is served by a randomly chosen node, which acts as a server/proxy server for this particular information object.

For each of the following experiments, a warm-up period of two hours is assumed ($\approx$ 75000 requests) during which the network is connected (*i.e.*, no failures) and the requests are processed and cached following the modularity-based caching algorithm (*"initialization period"*). We then proceed to simulate a disruptive scenario that perturbs the topology and a given number of f nodes fail. The performance of the proposed information resilience strategy is then monitored for a period of another two hours (*"observation period"*). This is assumed to be the time interval until either the "lost" nodes reestablish their connectivity or the network stabilizes and the routing tables (*i.e.*, FIBs) are repopulated.

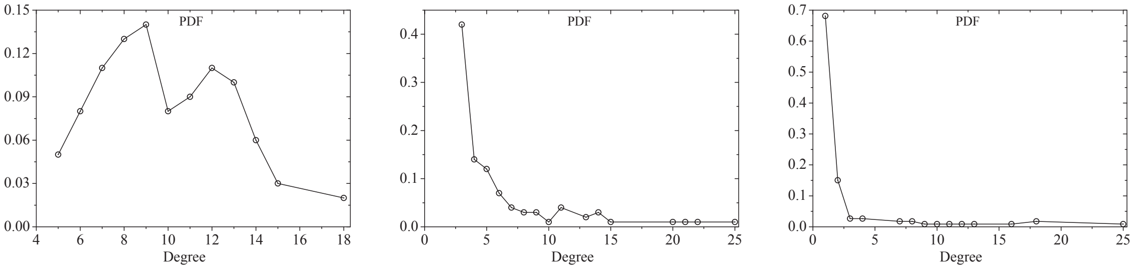
TABLE II: Summary of the evaluation parameter settings.

Parameter	Value
Number of information objects, M	10^5
Cache size, C	$0.01 \cdot M$
Popularity Zipf exponent, z	0.7
Failed nodes, f	$0.25 \cdot V$
Flooding scope, s	2

In the evaluation, the NDN's indiscriminate caching (*i.e.*, leave copy everywhere) with scoped flooding is used as the benchmark for performance comparison. This scheme is labeled as "*NDN+Scoped_flooding*" in the plots. We also used the spectral clustering algorithm [14] (labeled as "*Spectral*" in the plots) as an alternative partitioning scheme to further illustrate the effectiveness of the two used modularity-based algorithms. We simulated the proposed modularity-based caching, labeled as "*Newman_Girvan*" [9] and "*Newman_Eig*" [10] – *i.e.*, the same caching behavior is maintain throughout both initialization and observation periods. Further, for each of these schemes, we experimented with a variant whereby the routers neither cache nor evict content in their CS when the `Data` is returned via scoped flooding (these are labeled as "*Newman_Eig_No_Cache*", "*Newman_Girvan_No_Cache*" and "*Spectral_No_Cache*" respectively).

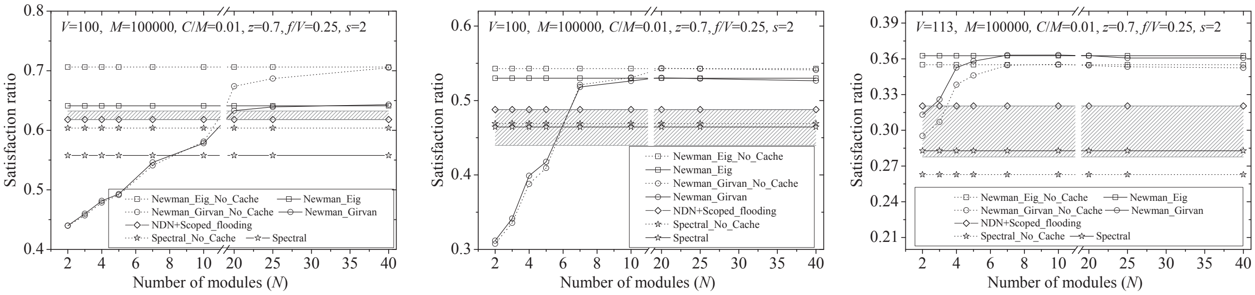
The leave copy everywhere approach of NDN [13] has been shown to be suboptimal, *e.g.*, [4], [23]. For this reason, the performance region (*i.e.*, gray pattern area in the plots) of various probabilistic caching schemes is also depicted, where each user caches a passing by object with some probability. This probability is varied in the space $[0.1, 1]$ (*i.e.*, using 10 different values), where a cache probability equal to 1 is the benchmark *NDN+Scoped_flooding* scheme. The examined probabilistic caching schemes is a way to differentiate the cached content of neighboring routers. Our rationale is to examine if such a simplistic scheme is capable to increase information resilience in perturbed ICNs, or the incorporation of modules can still increase resilience. Another approach to increase resilience would have been the deployment of CDN-like repositories throughout the network, where all the information objects would be replicated. However, this approach is host-centric and here we want to illustrate that a simple yet efficient assignment of nodes into communities can still increase information resilience without adding extra resources at the network. Definitely, this approach would further enhance the performance of the network during disruptions and is left for future investigation.

The evaluation is based on the satisfaction metric, *i.e.*, the percentage of `Interest` that have been satisfied (*i.e.*, found the requested object) during the observation period either at a cache of a router or at the corresponding server. Table II gives the default values for the various system parameters. For each setting, the simulation is repeated 20 times with random



(a) ER topology (b) SF topology (c) AT&T topology

Fig. 4: The degree distribution of the three topologies used in our simulations.



(a) ER topology (b) SF topology (c) AT&T topology

Fig. 5: The impact of the number of modules in the satisfaction ratio of the system.

failures and the average values are presented[4].

B. Impact of the number of communities/modules

For *Newman_Girvan*, we can indicate the number of modules N we want to create. This afforded the network operator with the capability to control the number of formed modules. However, unrealistic number of modules may force the algorithm to create modules that may not exhibit strong community characteristics. Thus, in this section, we evaluate the impact of this parameter on the different considered topologies. Fig. 5 shows the satisfaction ratios with different number of modules in the different topologies. Since the number of modules for the *Newman_Eig* and *Spectral* schemes cannot be controlled and there is no notion of modules in *NDN+Scoped_flooding*, they are plotted as straight lines for comparison.

For ER graph, since nodes are randomly connected, community structure only appears by chance and tightly knitted communities should not exist (theoretically). This is why, from Fig. 5(a), we see that at low number of modules, the satisfaction ratio achieved is actually lower than indiscriminate caching. However, when the number of modules is increased (*i.e.*, the mean size of each community decreases), the satisfaction ratio achieved is better. We attribute this effect to the reduced caching redundancy in the network since routers only cache content outside its modules. With power-law degree distribution (see Fig. 4(b)), hubs (nodes with high degrees) appear in scale-free graphs. This fosters the formation of communities since nodes are more likely to connect to these nodes with high connectivity, creating a skewed degree distribution. From Fig. 5(b), the satisfaction ratio achieved is already better than *NDN+Scoped_flooding* when the number of modules configured is $N > 6$. From Fig. 5(c), it is observed that the community effect is even stronger in the AT&T network (*i.e.*, satisfaction ratio exceeds that achieved by *NDN+Scoped_flooding* when $N > 3$). Finally, we note that using *Newman_Eig*, the resulting number of modules, N, is 16, 12 and 12 for ER, SF and AT&T network respectively (refer to Table I). Since the *Newman_Eig* always perform better than *Newman_Girvan* regardless of the number of modules that the latter forms, in the remainder of this section, *Newman_Girvan* is omitted and we only depict the performance of *Newman_Eig*.

From Fig. 5, is also observed that the different probabilistic caching schemes have a different impact in the satisfaction metric only when the degree distribution of the topologies follow a power law distribution (*i.e.*, SF and AT&T topologies). However, due to the usage of a small flooding scope, the leave copy everywhere approach still performs better than the rest of probabilistic schemes. In other words, the cache diversity accomplished by a probabilistic caching scheme is not enough to guarantee information resilience, since in a perturbed network the interests might not be able to travel far and the community characteristics should be incorporated in the caching decision process. Regarding the performance of the *Spectral* schemes, we observe that both variants perform significantly worse compared of the two modularity algorithms and also worse than the benchmark *NDN+Scoped_flooding*. This illustrates clearly the superiority of the other two modularity algorithms, and is an indication that a random or an inefficient clustering of the network is not sufficient to provide the required information resilience. We omit in the remainder

[4] For visualization clarity, the average values are depicted without the error intervals. However, we noticed only a $\pm 3\%$ variations on the performance of the 20 iterations from the average values.

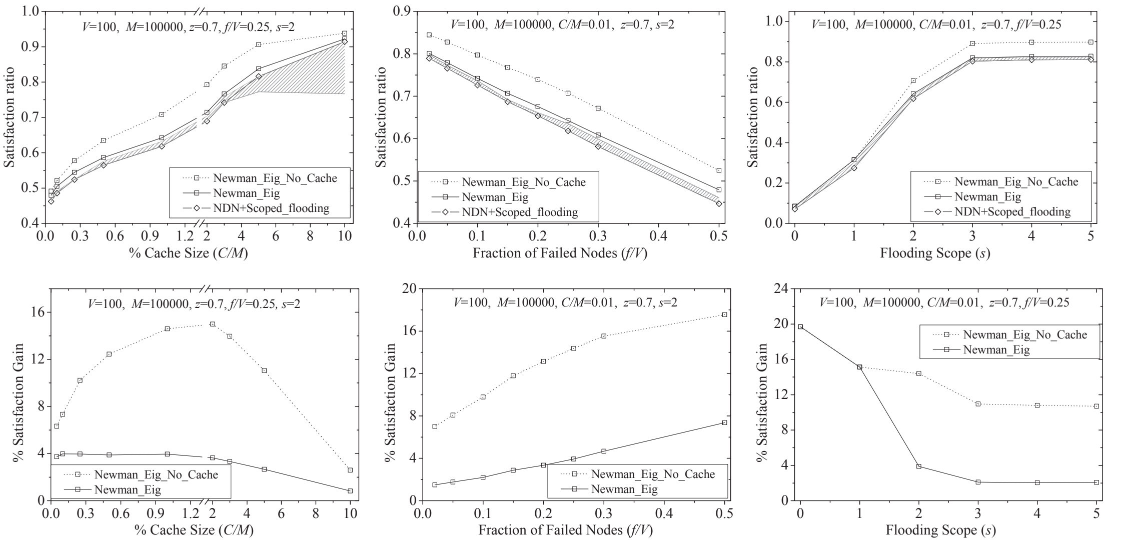


Fig. 6: The impact of the cache size, the ratio of the failed nodes and the scope of the `Interest` flooding in the satisfaction ratio at the ER topology.

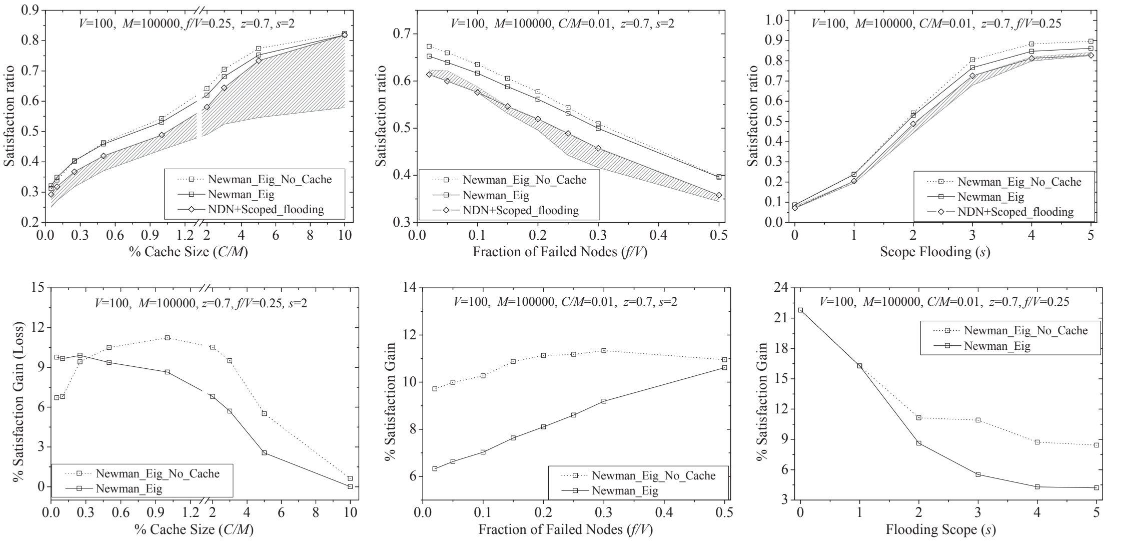


Fig. 7: The impact of the cache size, the ratio of the failed nodes and the scope of the `Interest` flooding in the satisfaction ratio at the SF topology.

of this section the *Spectral* schemes.

C. Impact of cache size

In this section, the impact of the cache size to the performance of the different resilience schemes is discussed. The first (left-most) columns in Fig. 6, Fig. 7 and Fig. 8 show the satisfaction ratio (top) and the ratio gain (bottom) compared against *NDN+Scoped_flooding* for ER, SF and AT&T networks respectively. In general, satisfaction ratio increases as the cache size increases for all schemes across all topologies, which is logical since there is more capacity to store information to satisfy interests. When benchmarked against *NDN+Scoped_flooding*, we observe that the gain achieved slowly decrease when larger cache size values are used. This is because the ability to cache large amount of information will naturally counter the effect of failures. If this is extrapolated to the extreme case where $C/M \approx 1$, then all schemes will have the same performance since all nodes can theoretically have a copy of all content in their CS. This is, of course, unrealistic as in real world, the cache size is generally extremely small compared to the content population in the Internet space. In such realistic range (*i.e.*, C/M small), we observe about 8% to 20% performance gain against *NDN+Scoped_flooding* for the different topologies.

D. Impact of the number of failed nodes

In this section, we evaluate the impact of failures in the network by introducing increasing number of failed nodes to the respective topologies under consideration. The second (middle) columns in Fig. 6, Fig. 7 and Fig. 8 show the

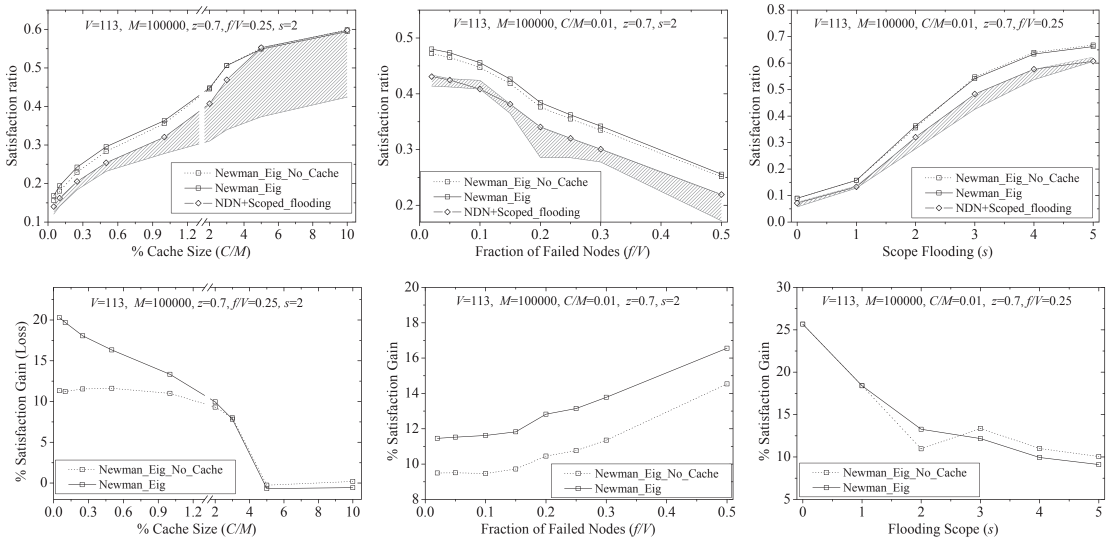


Fig. 8: The impact of the cache size, the ratio of the failed nodes and the scope of the `Interest` flooding in the satisfaction ratio at the AT&T topology.

satisfaction ratio (top) and the ratio gain (bottom) compared against *NDN+Scoped_flooding* for ER, SF and AT&T networks respectively. As expected, satisfaction ratio decreases as the number of failed nodes increases for all schemes across all topologies since more information sources are lost. However, under the same failure conditions, the *Newman_Eig* approaches consistently provide better satisfaction ratio. Furthermore, the worse the perturbations (*i.e.*, higher fraction of failed nodes), the higher the gain our approach achieves. This validates the original rationale that by caching selectively, based on community structure, we allow higher content diversity nearby and thus, promotes better cache hit rates. However, using only an probabilistic caching scheme, without incorporating the community structure is not enough to increase information resilience. From the gray pattern area in the corresponding plots we observe that the *Newman_Eig* variances always perform better than any probabilistic caching schemes. Note that this resilience property is also important in other network types (*e.g.*, cyber physical systems [24], [25]).

E. Impact of the flooding scope

In this section, we discuss the impact of the flooding scope to the performance of the resilience schemes. The third (right-most) column in Fig. 6, Fig. 7 and Fig. 8 shows the satisfaction ratio (top) and the ratio gain (bottom) compared against *NDN+Scoped_flooding* for ER, SF and AT&T networks respectively. In general, satisfaction ratio increases as the flooding scope, s, increases for all schemes across all topologies and saturates when s approaches the network diameter (*i.e.*, the gain decreases as s increases). This agrees with the finding in [18] where the probability of finding a cached copy of a content via scoped flooding decreases exponentially with the flooding radius. In addition, when the scoped flooding facility is disabled (*i.e.*, by setting $s = 0$), the modularity-based caching provides over 20% performance gain compared against *NDN+Scoped_flooding*. Enabling scoped flooding with increasing radius decreases this gain gradually but with the cost of high overhead. This is especially important since network diameters are usually small (*i.e.*, increasing s can quickly result in flooding the entire network).

F. Impact of different topologies

We notice that the gap of performance gain achieved between *Newman_Eig* and *Newman_Eig_No_Cache* across the three impact factors (namely cache size (left column), fraction of failed nodes (middle column) and flooding scope (right column) of Fig. 6, Fig. 7 and Fig. 8 respectively), in general, decreases from ER to SF to AT&T networks. This behaviour does not seem to depend on the degree distribution of the three topologies since while it is clear that ER does not share the same degree distribution with SF and AT&T networks, both SF and AT&T networks show, at least qualitatively, similar power-law distribution. As such, to gain further insights, we attempt to delve into the spectra of these networks.

In [26], [27], it has been found that Laplacian representation of graphs is better than adjacency matrix when used to compare the cospectrality of graphs. As such, we first compute the Laplacian matrices, Λ, for each topology (*i.e.*, $\Lambda = D - A$ where D is the degree matrix where the elements in the principal diagonal encodes the degree of each node and A is the adjacency matrix). Since the Laplacian matrices are still symmetric, we have real eigenvalues. The real value of all the eigenvalues (denoted by $\lambda_i; \forall i \in V$ and ordered as $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{V-1} \leq \lambda_V$) are plotted in Fig. 9. Also, since all the networks are connected (*i.e.*, only one giant component), the smallest eigenvalue is 0 with multiplicity 1.

From the figure, we surmise that having a low and "flatter" curve at the lower end of the spectrum fosters better gain for *Newman_Eig* than *Newman_Eig_No_Cache*. This conjecture follows the fact that relates to Cheegar's inequality [28] which stated that the sparsest cut of a graph can be approximated by the spectral gap of the graph. Subsequent to this, the Cheegar

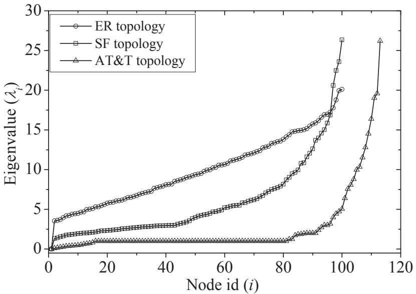


Fig. 9: The Laplacian eigenvalues of ER, SF and AT&T.

constant (sometimes known as the isoperimetric number) measures the existence of bottlenecks in a network. Based on this and from our results, ER is the least "bottlenecked" followed by SF and then lastly, AT&T. Following this, the results indicate that when the network has less bottlenecks, it is more beneficial to use the scheme with *No_Cache* variant and vice versa.

V. Conclusions

In this paper, we leverage the central premises of ICN and investigate the provision of information resilience through modularity-based caching in networks under perturbations (*i.e.*, network with node/link failures). This work provides a first stab on the problem of providing information resilience. In our approach, each caching network element makes its caching decision based on the community it belongs to as well as the origin community of the content. The proposed strategy, ensures high information diversity within each community and as such, enables many more distinct content to be cached near each node. Extensive simulations were conducted based on both synthetic and real network topologies and the impact of different parameters (*e.g.*, cache size, number of failures, flooding scope, *etc.*) to the performance of the proposed approach was investigated. The results indicate significant improvement in terms of request satisfaction ratio when compared with non-community-aware caching approaches. In fact, we find that with the proposed approach, the more severe the network is perturbed, the higher the satisfaction ratio gain is achieved. In addition, the results also show that the choice of community detection algorithm is important. In particular, it is observed that modularity-based algorithms outperform the traditional spectral clustering approach based on graph cut size.

Acknowledgment

This work has been supported by the EC H2020 ICN2020 project (GA no. 723014) and the CHISTERA/EPSRC UK project CONCERT (GA no. EP/L018535/1).

References

[1] G. Xylomenos, C. Ververidis, V. Siris, N. Fotiou, C. Tsilopoulos, X. Vasilakos, K. Katsaros, and G. Polyzos, "A survey of information-centric networking research," *IEEE Communications Surveys Tutorials*, vol. 16, no. 2, pp. 1024–1049, 2014.

[2] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A survey of information-centric networking," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 26–36, July 2012.

[3] V. Sourlas, L. Tassiulas, I. Psaras, and G. Pavlou, "Information resilience through user-assisted caching in disruptive content-centric networks," in *IFIP Networking*, May 2015, pp. 1–9.

[4] W. K. Chai, D. He, I. Psaras, and G. Pavlou, "Cache less for more in information-centric networks (extended version)," *Computer Communications*, vol. 36, no. 7, pp. 758 – 770, 2013.

[5] M. Zhang, H. Luo, and H. Zhang, "A survey of caching mechanisms in information-centric networking," *IEEE Communications Surveys Tutorials*, vol. 17, no. 3, pp. 1473–1499, 2015.

[6] P. Smith, D. Hutchison, J. Sterbenz, M. Schller, A. Fessi, M. Karaliopoulos, C. Lac, and B. Plattner, "Network resilience: a systematic approach," *IEEE Communications Magazine*, vol. 49, no. 7, pp. 88–97, July 2011.

[7] S. Trajanovski, J. Martn-Hernndez, W. Winterbach, and P. Van Mieghem, "Robustness envelopes of networks," *Journal of Complex Networks*, vol. 1, no. 1, 2013.

[8] M. F. Al-Naday, M. J. Reed, D. Trossen, and K. Yang, "Information resilience: source recovery in an information-centric network," *IEEE Network*, vol. 28, no. 3, pp. 36–42, May 2014.

[9] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, Feb 2004.

[10] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical review E*, vol. 74, no. 3, 2006.

[11] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 35, pp. 75 – 174, 2010.

[12] U. Brandes, D. Delling, M. Gaertler, R. Grke, M. Hoefer, Z. Nikoloski, and D. Wagner, "On Modularity – NP-Completeness and Beyond," 2006.

[13] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *ACM CoNEXT*, 2009, pp. 1–12.

[14] M. Fiedler, "Algebraic connectivity of graphs," *Czech. Math. J.*, vol. 23, no. 98, pp. 298–305, 1973.

[15] T. Koponen, M. Chawla, B.-G. Chun, A. Ermolinskiy, K. H. Kim, S. Shenker, and I. Stoica, "A data-oriented (and beyond) network architecture," in *ACM SIGCOMM*, 2007, pp. 181–192.

[16] W. K. Chai, N. Wang, I. Psaras, G. Pavlou, C. Wang, G. de Blas, F. Salguero, L. Liang, S. Spirou, A. Beben, and E. Hadjioannou, "Curling: Content-ubiquitous resolution and delivery infrastructure for next-generation services," *Comm. Mag.*, vol. 49, no. 3, 2011.

[17] G. Pavlou, N. Wang, W. K. Chai, and I. Psaras, "Internet-scale content mediation in information-centric networks," *Annals of Telecommunications*, vol. 68, no. 3, pp. 167–177, 2013.

[18] L. Wang, S. Bayhan, J. Ott, J. Kangasharju, A. Sathiaseelan, and J. Crowcroft, "Pro-diluvian: Understanding scoped-flooding for content discovery in information-centric networking," in *ACM International Conference on ICN*, 2015, pp. 9–18.

[19] A. K. M. M. Hoque, S. O. Amin, A. Alyyan, B. Zhang, L. Zhang, and L. Wang, "Nlsr: Named-data link state routing protocol," in *ACM SIGCOMM Workshop on ICN*, 2013, pp. 15–20.

[20] P. Erdös and A. Rényi, "On random graphs, I," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959.

[21] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[22] N. Spring, R. Mahajan, and D. Wetherall, "Measuring isp topologies with rocketfuel," in *ACM SIGCOMM*, 2002, pp. 133–145.

[23] I. Psaras, W. K. Chai, and G. Pavlou, "In-network cache management and resource allocation for information-centric networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 11, pp. 2920–2931, Nov. 2014.

[24] W. K. Chai, N. Wang, K. V. Katsaros, G. Kamel, G. Pavlou, S. Melis, M. Hoefling, B. Vieira, P. Romano, S. Sarri, T. Tesfay, B. Yang, F. Heimgaertner, M. Pignati, M. Paolone, M. Menth, E. Poll, M. Mampaey, H. Bontius, and C. Develder, "An information-centric communication infrastructure for real-time state estimation of active distribution networks," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 2134–2146, 2015.

[25] K. V. Katsaros, W. K. Chai, N. Wang, G. Pavlou, H. Bontius, and M. Paolone, "Information-centric networking for machine-to-machine data delivery: a case study in smart grid applications," *IEEE Network*, vol. 28, no. 3, pp. 58–64, May 2014.

[26] R. C. Wilson and P. Zhu, "A study of graph spectra for comparing graphs and trees," *Pattern Recogn.*, vol. 41, no. 9, pp. 2833–2841, Sep. 2008.

[27] W. H. Haemers and E. Spence, "Enumeration of cospectral graphs," *European Journal of Combinatorics*, vol. 25, no. 2, pp. 199 – 211, 2004.

[28] J. Cheeger, *A lower bound for the smallest eigenvalue of the laplacian*, 1970.