Efficient Data Forwarding in Mobile Social Networks with Diverse Connectivity Characteristics

Xiaomei Zhang and Guohong Cao
Department of Computer Science and Engineering
The Pennsylvania State University, University Park, PA, 16802
Email: {xqz5057,gcao}@cse.psu.edu

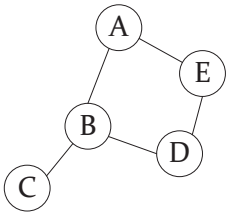
***Abstract*—Mobile Social Network (MSN) with diverse connectivity characteristics is a combination of opportunistic network and mobile ad hoc network. Since the major difficulty of data forwarding is the opportunistic part, techniques designed for opportunistic networks are commonly used to forward data in MSNs. However, this may not be the best solution since they do not consider the ubiquitous existences of *Transient Connected Components (TCCs)*, where nodes inside a TCC can reach each other by multi-hop wireless communications. In this paper, we first identify the existence of TCCs and analyze their properties based on five real traces. Then, we propose TCC-aware data forwarding strategies which exploit the special characteristics of TCCs to increase the contact opportunities and then improve the performance of data forwarding. Trace-driven simulations show that our TCC-aware data forwarding strategies outperform existing data forwarding strategies in terms of data delivery ratio and network overhead.**

I. Introduction

In Mobile Social Networks (MSNs), human-carried mobile devices opportunistically form wireless peer-to-peer connections with each other in the absence of the network infrastructure [1]. Due to the unpredictable human mobility, it is hard to maintain end-to-end connections. As a result, "carry-and-forward" [2][3] is used, where mobile nodes physically "carry" the data, and forward the data when contacting a node with "higher" forwarding capability.

Although most research assumes that MSNs are highly sparse, a close scrutiny on most current MSN traces reveals that the connectivity inside a network is diverse, and there are ubiquitous existences of *Transient Connected Components (TCCs)*. Inside TCCs, nodes have transient contacts with each other and form a connected component. For example, students in a classroom have transient connections with each other, and vehicles on highways form platoons and have transient connections inside a platoon [4].

An MSN with diverse connectivity characteristics is a combination of opportunistic network and mobile ad hoc network (MANET). Inside a TCC, a MANET is formed where nodes can reach each other by multi-hop wireless communications. Outside of TCC, nodes contact each other opportunistically through "carry-and-forward". Since the major difficulty of data forwarding is the opportunistic part, techniques designed for opportunistic networks are commonly used to forward data in MSNs. However, this may not be the best solution since they do not consider the diverse connectivity characteristics of MSNs. For example, Figure 1 (a) shows a TCC of five nodes and Figure 1 (b) shows their forwarding metrics which represent the capability of forwarding data to other nodes (e.g., node centrality [5]). Node A has the data which will be forwarded to the destination through "carry-and-forward". Based on existing techniques, the data should be forwarded to a contacted node with higher forwarding metric. In this example, node A has two possible contacts: B and E (contacts are represented by lines). Since node B has lower forwarding metric (2.1) than A's (3.0), B will not get the data. E has higher forwarding metric (4.5) than A, and thus A forwards the data to E. After E receives the data, its contact D, which has higher forwarding metric (5.8), will get the data. D will not forward the data to its contact B which has a lower forwarding metric. Although C has a much higher forwarding metric (7.0) and much higher chance of reaching the destination, the data will not be forwarded to C since it is not the contact of D (i.e., no line between them). However, since C and A are within the same TCC, it is better for them to exchange their forwarding metrics through multi-hop wireless communication, and then it will be possible for C to forward the data to the destination.



(a)

Nodes	Forwarding metric
A	3.0
B	2.1
C	7.0
D	5.8
E	4.5

(b)

Figure 1. The left figure shows a TCC of five nodes. The right table shows their forwarding metrics. A line between two nodes means that they are within communication distance.

In this paper, we address the problem of data forwarding in MSNs with diverse connectivity characteristics by proposing two TCC-aware data forwarding strategies. Since a TCC

This work was supported in part by Network Science CTA under grant W911NF-09-2-0053.

is a MANET, it is treated as one component and data carriers in this TCC are selected for opportunistic data forwarding. More specifically, the paper has two contributions.

1) We identify the existence of TCCs and analyze their properties based on five traces. We find that there are significant number of TCCs in MSNs, and the distributions of TCC size and node degree follow exponential distribution. By treating multi-hop wireless communications inside TCCs as indirect contacts, through theoretical analyses, we show that the contact opportunities can be significantly increased in all traces.
2) We first propose a TCC-aware data forwarding strategy to exploit TCCs to improve the performance of data forwarding in MSNs. In our solution, nodes inside a TCC exchange their forwarding metrics through multi-hop wireless communications, and the node with the highest forwarding metric is selected to get a replicated data copy. Although the TCC-aware data forwarding strategy can increase the data delivery ratio, it increases the data copies in the network. To address this problem, we enhance the TCC-aware data forwarding by selecting an optimal set of nodes in the TCC to avoid overlap in their contacts and maximize the data forwarding opportunity with a small number of nodes. Trace-driven simulations show that our TCC-aware data forwarding strategies outperform existing data forwarding strategies with less network overhead.

The rest of the paper is organized as follows. In Section II, we identify the properties of TCCs based on five traces. Section III presents our TCC-aware data forwarding strategies. In Section IV, we evaluate the performance of the TCC-aware data forwarding strategies. Section V reviews related work, and Section VI concludes the paper.

II. Trace-based TCC analysis

In this section, we identify the existence of TCCs and analyze their properties based on five realistic MSN traces.

A. Traces

We study the properties of TCCs based on five traces: *Social Evolution* [6], *Friends & Family* [7], *Reality Mining* [8], *Infocom* [9] and *UCSD* [10]. These traces record contacts among users carrying different kinds of mobile devices. The first three traces are collected by the MIT Reality group based on Bluetooth on smartphones. Among them, *Social Evolution* records the contacts among students in an undergraduate dormitory. *Friends & Family* records the contacts among members of a young-family residential community. *Reality Mining* tracks the contacts between individuals in research labs. The trace of *Infocom* is collected in a conference environment by recording the contacts between conference attendants carrying imotes. In these four traces, mobile devices periodically detect their peers via Bluetooth interfaces. A contact is recorded when two mobile devices move into the detection range of each other. The *UCSD* trace is collected at a campus scale, where the devices are WiFi enabled PDAs. These devices search for nearby WiFi Access Points (APs), and a contact is detected when two devices detect the same AP. The details of these five traces are shown in Table 1.

B. Properties of the Contact Graph

Based on the collected traces, we can draw contact graphs which consist of mobile devices (nodes) and their contacts (edges), with which we study the transient contact status between nodes. A contact graph can be drawn at each time point, i.e., there is an edge between two nodes if they are contacting each other at that time point. Here, a contact graph is drawn every 10 minutes in the traces.

The extracted contact graphs have some interesting properties, and one of them is related to the distribution of the node degree. The Complementary Cumulative Distribution Function (CCDF) of the node degree $P(K > k)$, which represents the probability that a node has more than k contacts with other nodes, follows exponential distribution for $k \geq 1$:

$$P(K > k) = e^{-\frac{k}{k^*}} \qquad k \geq 1$$

where k^* is the exponential constant. By recognizing that $e^{-\frac{k}{k^*}} = 1$ when $k = 0$, we have the following:

$$P(K > k) = e^{-\frac{k}{k^*}} \qquad k \geq 0 \tag{1}$$

The plot of CCDF $P(K > k)$ is shown in Figure 2. The CCDF curve is drawn in a semi-log plot (*log* scale on y axis), where the exponential distribution is a straight line. The Least Square approach [11] is applied to fit the degree distribution with the exponential distribution. As we can see in Figure 2, the degree distributions can be well approximated by exponential distributions in most traces.

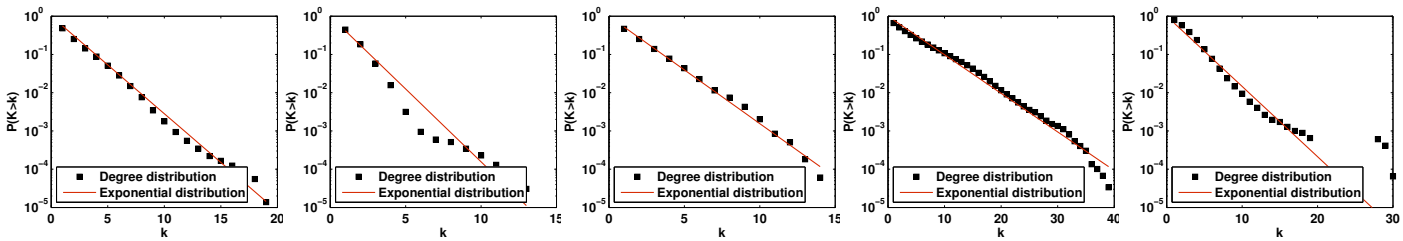
Since the contact graphs satisfy exponential degree distributions, most nodes have low degrees. This is because the contact graph is partitioned into connected components, where nodes are only connected to some nodes in the same connected component, so that most nodes have low degrees. For example, in the trace of *Friends & Family*, people in the residence community form small connected components when they stay with their families. In the trace of *Reality Mining*, students form connected components when they are in the same laboratory. Since the connected components only appear temporarily, they are referred to as *Transient Connected Components (TCCs)*.

C. Properties of TCC

We study the properties of TCCs in this subsection. The first property is related to *Giant Connected Component (GCC)*, which is the largest connected component of the

Table I
TRACE SUMMARY

Trace	Social Evolution	Friends & Family	Reality Mining	Infocom	UCSD
Device	Cell Phone	Cell Phone	Cell Phone	iMote	WiFi
Network type	Bluetooth	Bluetooth	Bluetooth	Bluetooth	WiFi
Number of devices	80	84	97	98	275
Number of contacts	200,841	47,774	78,098	126,804	194,681
Start Date	2008-10-01	2010-11-01	2004-09-27	2006-04-23	2002-09-22
Durations(days)	243	142	246	4	776
Granularity(secs)	300	300	300	120	20



(a) Social Evolution ($k^* = 1.70$) (b) Friends & Family ($k^* = 1.14$) (c) Reality Mining ($k^* = 1.55$) (d) Infocom ($k^* = 4.31$) (e) UCSD ($k^* = 2.36$)

Figure 2. The distribution of node degree can be approximated by exponential distribution. k^* is the exponential constant.

contact graph. The GCC is commonly used to represent the overall connectivity of the network [12], and the property of the GCC is closely related to the degree distribution [13][14][15]. Based on [15], in a random graph with exponential degree distribution, if the exponential constant k^* is larger than 1, there exists a GCC with size that scales linearly with the size of graph. With the increase of k^*, the size of the GCC also increases. Since k^* in all traces is larger than 1 as shown in Figure 2, we can expect that there are large GCCs inside these contact graphs.

Figure 3 uses *boxplot* to show the size of the largest TCC as a proportion of the network size, where the largest TCC is detected every 10 minutes. Here, the network size is the number of nodes that are in contact with some nodes in the trace. Other nodes may turn off their devices or contact nodes that are not included in the trace, and these nodes are not included in our analysis. The results in Figure 3 are consistent with our conclusion that a larger exponential constant k^* leads to a larger GCC. For example, the *Family & Friends* trace has the smallest GCC, because k^* is only slightly larger than 1. For the trace of *Infocom*, k^* is larger than 4 and thus most nodes belong to the GCC.

Even though the size of GCC implies the overall connectivity of the network, it is not enough to characterize the complete TCC structure. Therefore, we are also interested in the distribution of the TCC size. Figure 4 plots the distribution of the TCC size. Similar to the degree distribution, the TCC size distribution also follows exponential distribution. The CCDF of the TCC size $P(S > s)$ is exponential for

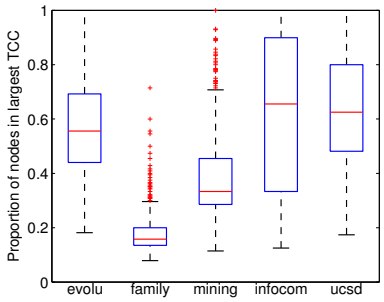


Figure 3. The boxplot of the largest TCC size as a proportion of the network size. The middle line inside the box represents the median. The lower and upper edges of the box represents the 25th and 75th percentiles, respectively.

$s \geq 2$:

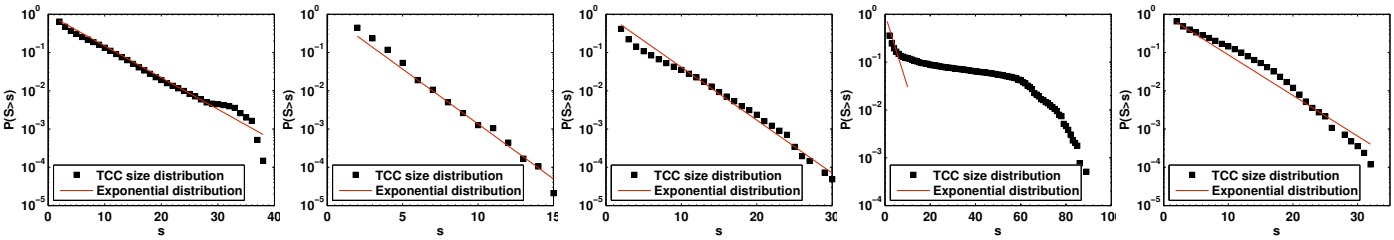
$$P(S > s) = \begin{cases} e^{-\frac{s}{s^*}} & s \geq 2 \\ 1 & 0 \leq s \leq 1 \end{cases} \quad (2)$$

where s^* is the exponential constant.

In the *Infocom* trace, it only follows exponential distribution when the TCC size is small, i.e., in the range of $[2, 10]$. This is due to the fact that the network always consists of a GCC which includes most of the nodes in the trace, and then the remaining nodes only form small TCCs.

D. Increasing Contacts with TCC

Inside a TCC, nodes can communicate with each other through multi-hop wireless communication, which is much faster than "carry-and-forward". Therefore, as long as two



(a) Social Evolution ($s^* = 5.25$) (b) Friends & Family ($s^* = 1.52$) (c) Reality Mining ($s^* = 3.13$) (d) Infocom ($s^* = 2.86$) (e) UCSD ($s^* = 4.09$)

Figure 4. The distribution of TCC size can be well approximated by exponential distribution for most of the traces. s^* is the exponential constant.

nodes are within a TCC, they have *TCC-contact*, which can be *direct contact* or *indirect contact*. *Direct contacts* are contacts between two nodes within one-hop communication distance, and *indirect contacts* are contacts between nodes within the same TCC but multi-hop away. Next, we analyze whether and how contact opportunities are increased by considering TCC-contacts. We find that the increase of contact opportunities is related to the distributions of node degree and TCC size. Specifically, we have the following theorem:

***Theorem* 1.** *For a contact graph, the contact opportunities can be increased by considering TCC-contacts, if the following two conditions are satisfied:*

1) *The distribution of node degree and the distribution of TCC size both follow exponential distributions with exponential constants k^* and s^* respectively. (Equations (1) and (2))*
2) *$k^* < \hat{k}^*$, where $\hat{k}^*$ is a function of s^*:*

$$\hat{k}^* = \frac{1}{\log \frac{2\cdot(e^{-\frac{1}{s^*}}+(1-e^{-\frac{1}{s^*}})^3)}{3e^{-\frac{1}{s^*}}-1+(1-e^{-\frac{1}{s^*}})^3}} \tag{3}$$

The proof of Theorem 1 can be found in Appendix A. In the proof, we also compute the ratio of TCC-contacts to direct contacts, which can be determined by two exponential constants k^* and s^*:

$$\frac{m_t}{m_d} = 2(1-e^{-\frac{1}{k^*}}) \cdot \frac{e^{-\frac{1}{s^*}} + (1-e^{-\frac{1}{s^*}})^3}{1-e^{-\frac{1}{s^*}} + (1-e^{-\frac{1}{s^*}})^3} \tag{4}$$

where m_t is the number of TCC-contacts and m_d is the number of direct contacts. With Equation (4), we can estimate the increased contact opportunities as long as the node degree and TCC size follow exponential distributions.

Table II lists k^*, s^* and the value of $\hat{k}^*$ in terms of s^* in four traces. (Since TCC size distribution for *Infocom* trace does not follow exponential distribution, we do not include it in Table II.) We also did experiments to verify if the estimated value $\frac{m_t}{m_d}$ in Equation (4) (i.e., Est. $\frac{m_t}{m_d}$) is consistent with the actual value (i.e., Act. $\frac{m_t}{m_d}$) and the results are shown in Table II. As can be seen, k^* is smaller than $\hat{k}^*$ in all traces, which indicates that the contact opportunities are increased in all these traces. We also find that the error of the estimated $\frac{m_t}{m_d}$ compared to the actual value is very small, which proves the accuracy of our estimation. Thus, with the distributions of node degree and TCC size, we can accurately estimate the increase of contact opportunities.

We also did experiment on the *Infocom* trace and find the $\frac{m_t}{m_d}$ is 8.86, which indicates that the contact opportunities are also significantly increased in the *Infocom* trace.

Table II
TCC-CONTACTS & DIRECT CONTACTS

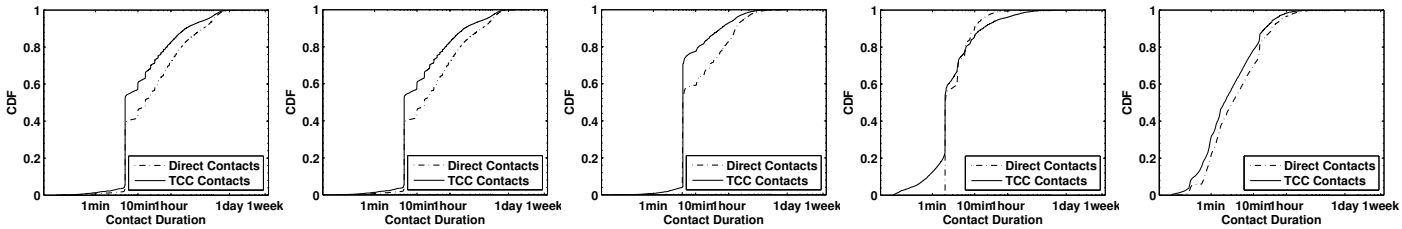
Trace	k^*	s^*	$\hat{k}^*$	Est. $\frac{m_t}{m_d}$	Act. $\frac{m_t}{m_d}$
Social Evolution	1.70	5.25	8.80	4.14	4.33
Friends & Family	1.14	1.52	1.57	1.24	1.38
Reality Mining	1.55	3.13	4.56	2.42	2.00
UCSD	2.36	4.09	6.47	2.41	2.31

E. Durations of TCC-contacts

Because TCCs are formed when nodes have direct contacts with each other, the duration of the TCC-contact is smaller than the duration of direct contact. The comparisons between the durations of TCC-contacts and direct contacts in five traces are shown in Figure 5. As shown in the figure, even though TCC-contacts have smaller durations than direct contacts, the median durations for TCC-contacts are still several minutes. Therefore, by considering both direct and indirect contacts inside TCCs, contact opportunities are increased and the contact durations are still pretty long. This property can be exploited to design more efficient data forwarding strategies, as shown in the next section.

III. TCC-AWARE DATA FORWARDING STRATEGIES

Data forwarding in MSNs is difficult due to the opportunistic nature of the network. Being aware of the existence of TCCs, we propose better data forwarding strategies by utilizing the contact opportunities in TCCs. Inside a TCC, nodes can reach each other by multi-hop wireless communications, which will significantly increase the contact opportunities and increase the chance of forwarding data to the destination. In this section, we first present our



(a) Social Evolution (b) Friends & Family (c) Reality Mining (d) Infocom (e) UCSD

Figure 5. The durations of direct contacts and TCC-contacts.

TCC-aware data forwarding strategy and then improve its performance with an enhanced version.

A. TCC-Aware Data Forwarding Strategy

The objective of the TCC-aware data forwarding strategy is to utilize TCCs to improve the performance of data forwarding in MSNs.

1) Identifying the TCC: To utilize TCCs, the first step is to identify the TCCs. A node can detect the nodes in the current TCC by broadcasting inside the TCC, and each node receiving the broadcast sends an acknowledgement. In order to know what data items are being forwarded in the current TCC, the node receiving the broadcast message replies with an acknowledgement that carries the information about the data items in its buffer. By collecting acknowledgements from other nodes in the TCC, the original sender can identify the nodes and the forwarded data in its current TCC.

2) Forwarding Strategy: For each data item created in the network, we intend to forward it from the source to the destination within the time constraint T. In the process of forwarding, data can be replicated and forwarded to other nodes in accordance with specific strategies. The forwarding is successful as long as one data copy of this data item arrives at the destination.

Strategies designed for opportunistic networks are commonly used to forward data in MSNs. These data forwarding strategies are normally based on social-aware forwarding metrics such as centrality [3][16][5], which quantify mobile node's capability of contacting others in the network. When a node carrying data contacts another node, the data packet is forwarded based on *Compare-and-Forward* [3][2]; i.e., the data carrier forwards the data if and only if the contacted node has a higher centrality and does not have this data. The original data carrier also keeps a copy after forwarding the data.

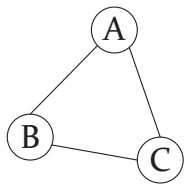
In our TCC-aware data forwarding strategy, a similar method is used. However, the data forwarding decisions are not limited to these two contacted nodes, but among all nodes in the TCC. A TCC can be formed by merging two existing TCCs or by adding new nodes to existing TCCs. For example, nodes A and B are originally from two different TCCs before they contact. After A contacts B, the two TCCs are merged to a new TCC.

Algorithm 1 TCC-aware Data Forwarding Strategy when two Nodes A, B Contact

$\mathcal{M}_A \leftarrow$ the set of nodes in A's original TCC
$\mathcal{M}_B \leftarrow$ the set of nodes in B's original TCC
if $\mathcal{M}_A = \mathcal{M}_B$ **then**
  Do NOTHING /*A, B *are already in the same TCC* */
else
  /**Forwarding decision inside the TCC will be made centrally at a command node C, which is chosen from A, B.**/
  if $|\mathcal{M}_A| \leq |\mathcal{M}_B|$ **then**
    $C \leftarrow B$
  else
    $C \leftarrow A$
  end if
  /**C will identify the nodes inside the new TCC and the data items carried by them, and make the forwarding decision.**/
  $\mathcal{M} \leftarrow \mathcal{M}_A \bigcup \mathcal{M}_B$
  $\mathcal{D} \leftarrow$ the set of unique data items in the new TCC
  for each data item $d \in D$ **do**
    if $\mathcal{M}$ includes d's destination **then**
      d is forwarded to d's destination
      Go to next data item
    end if
    $H_d \leftarrow$ node with the highest centrality
    if H_d has the data **then**
      Do NOTHING and go to the next data item
    else
      d is forwarded to H_d
    end if
  end for
end if

When two nodes contact, they first check if they are already in the same TCC before the contact. They detect their original TCCs by using broadcast messages described in Section III-A1. As long as one of them receives acknowledgement from the other one, they are already within the same TCC. If the detected TCCs are different for the two nodes, these two TCCs are merged when they contact.

After a new TCC is formed, a simple compare-and-forward strategy is used to forward data to nodes with higher centrality. However, this strategy creates many data copies. To reduce redundancy, data is only forwarded to the node with the highest centrality, and then at most one additional data copy is created in a TCC.



(a)

	A	B	C	D	E	F	G
A	x	1	1	0.7	0.6	0.3	0.2
B	1	x	1	0.1	0.1	0.6	0.5
C	1	1	x	0.5	0.6	0.1	0.1

(b)

Figure 6. The left figure shows a TCC of three nodes. The right table shows their contact probabilities with all other nodes in the network. The network has 7 nodes.

3) The Algorithm: The whole process of the TCC-aware data forwarding strategy is outlined in Algorithm 1. When two nodes contact, they first check if they belong to the same TCC (Lines $3 \sim 5$). If so, the data packet has already been forwarded in the TCC. If not, a new TCC is formed, and the data items will be forwarded inside the new TCC.

The forwarding decision inside this TCC is made centrally at a node, denoted as "command" node, which is chosen from the two nodes in contact. The node that has more nodes in its original TCC will be the command node (Lines $6 \sim 10$). Then, the command node checks all the nodes and their carried data items in the new TCC and makes the forwarding decisions (Lines $11 \sim 24$). Specifically, for each data item that exists in the TCC, if its destination is inside this TCC, it is directly forwarded to the destination (Lines $14 \sim 17$). Otherwise, the command node checks if the node with the highest centrality has the data. If the node has the data, nothing needs to be done. Otherwise, one data copy is created and forwarded to the node with the highest centrality (Lines $18 \sim 23$).

B. Enhanced TCC-aware Data Forwarding Strategy

1) Motivation: In the TCC-aware data forwarding strategy, the node with the highest centrality in the TCC gets one data copy and the original data carriers also keep their data copies. However, this may not be the best option in many cases. For example, Figure 6 (a) shows a TCC of three nodes and Figure 6 (b) lists each node's contact probability with others in the network. If the centrality metric is based on the Cumulative Contacting Probability (CCP) [5], the CCP of node A is $1 + 1 + 0.7 + 0.6 + 0.3 + 0.2 = 3.8$, the CCP of node B is $1 + 1 + 0.1 + 0.1 + 0.6 + 0.5 = 3.3$, and the CCP of node C is $1 + 1 + 0.5 + 0.6 + 0.1 + 0.1 = 3.3$. If C originally carries the data, A will receive one data copy since it has the highest centrality. However, nodes A and C have similar contact probabilities to other nodes B, D, E, F and G; i.e., they both have high contact probabilities with B, D and E, and low contact probabilities with F and G. Thus, keeping data copies at A and C may not help too much since their contact capabilities to other nodes have a large "overlap".

To deal with this problem, it is better to choose A and B as data carriers, because B has high contact probabilities to F and G, which is complement to A. Even though B and C have the same CCP centrality, using A and B as data carriers is better than using A and C. Thus, we propose an enhanced TCC-aware data forwarding strategy, where data carriers are selected to maximize the data forwarding opportunity. Different from the previous TCC-aware data forwarding strategy, which selects the highest centrality node as the data carrier, we select a set of nodes as data carriers which can maximize the data forwarding opportunity.

2) Set Centrality: We first define the concept of *set centrality* to quantify the data forwarding capability of a set of nodes. Given that a data item is created at time 0, and expires at time T, the *set centrality* of a node set $\mathcal{S}$ at time t is defined as follows:

Definition* 1.** *The* ***set centrality *of a node set $\mathcal{S}$ at time $t < T$ is defined as the summation of their overall probability to contact each of the remaining nodes in $\mathcal{N} \setminus \mathcal{S}$ before time T.*

$$C_{\mathcal{S}}(t) = \sum_{i \in \mathcal{N} \setminus \mathcal{S}} (1 - \prod_{j \in \mathcal{S}} (1 - p_{ji}(T - t))) \tag{5}$$

where $\mathcal{N}$ is the set of nodes in the network, and $p_{ji}(T - t)$ is the probability that node j will contact node i within time $T - t$.

Assume symmetric contacts between nodes i and j, and then we have $p_{ij}(T-t) = p_{ji}(T-t)$. Since the inter-contact time between node i and node j has been experimentally validated in [5] to follow an exponential distribution with rate parameter λ_{ij}, the probability that the two nodes will contact before T can be calculated as:

$$p_{ij}(T - t) = p_{ji}(T - t) = 1 - e^{-\lambda_{ij}(T-t)} \tag{6}$$

3) Selecting the Optimal Set of Data Carriers: We assume that there are k data copies carried by nodes in a TCC (denoted as $\mathcal{M}$), where $k < |\mathcal{M}|$. Our objective is to choose an optimal node set $\mathcal{S}^*$ of size k with the highest set centrality, where $\mathcal{S}^* \subset \mathcal{M}$. Nodes in the optimal node set $\mathcal{S}^*$ will be the new data carriers for the k data copies. The detection of the optimal node set $\mathcal{S}^*$ can be formalized as an optimization problem:

$$\max \sum_{i \in \mathcal{N}} (1 - x_i)(1 - \prod_{j \in \mathcal{N}} (1 - x_j \cdot p_{ji}(T - t))) \tag{7}$$

$$s.t. \quad x_i \in \{0, 1\}, \qquad \forall i \in \mathcal{M} \tag{8}$$

$$x_i = 0, \qquad \forall i \in \mathcal{N} \setminus \mathcal{M} \tag{9}$$

$$\sum_{i \in \mathcal{M}} x_i = k \tag{10}$$

where $x_i \in \{0, 1\}$ indicates whether the node i is selected to the optimal node set $\mathcal{S}^*$. Formula (7) maximizes the set centrality of the selected nodes. Since $\mathcal{S}^*$ is selected

Algorithm 2 Enhanced TCC-aware Data Forwarding Strategy when two Nodes A, B Contact

1: $\mathcal{M}_A \leftarrow$ the set of nodes in A's original TCC
2: $\mathcal{M}_B \leftarrow$ the set of nodes in B's original TCC
3: **if** $\mathcal{M}_A = \mathcal{M}_B$ **then**
4: Do NOTHING /*A, B *are already in the same TCC* */
5: **else**
/* *Forwarding decision inside the TCC will be made centrally at a command node C, which is chosen from A, B.**/
6: **if** $|\mathcal{M}_A| \leq |\mathcal{M}_B|$ **then**
7: $C \leftarrow B$
8: **else**
9: $C \leftarrow A$
10: **end if**
/*C *will identify the nodes inside the new TCC and the data items carried by them, and make the forwarding decision.**/
11: $\mathcal{M} \leftarrow \mathcal{M}_A \bigcup \mathcal{M}_B$
12: $\mathcal{D} \leftarrow$ the set of unique data items in the new TCC
13: **for** each data item $d \in D$ **do**
14: **if** $\mathcal{M}$ includes d's destination **then**
15: d is forwarded to d's destination
16: Go to next data item
17: **end if**
18: $k_d \leftarrow$ number of copies of d in $\mathcal{M}$
19: $H_d \leftarrow$ node with the highest centrality
20: **if** H_d does not have d **then**
21: $k_d \leftarrow k_d + 1$ /**Add one extra data copy**/
22: **end if**
23: $\mathcal{S}^* \leftarrow$ the optimal set of k_d nodes
/**Decide $\mathcal{S}^*$ as discussed in Section III-B3.**/
/**Nodes in $\mathcal{S}^*$ will be new carries of data item d.**/
24: Data are forwarded from old carriers to nodes in $\mathcal{S}^*$
25: **end for**
26: **end if**

from $\mathcal{M}$, $x_i \in \{0, 1\}$ for nodes inside $\mathcal{M}$, and $x_i = 0$ for nodes outside of $\mathcal{M}$ as depicted in (8) and (9) respectively. Formula (10) indicates that the number of selected nodes is k.

The optimization problem can be easily solved using dynamic programming similar to the *knapsack problem* [17]. With dynamic programming, the optimization problem can be solved with time complexity $O(|\mathcal{M}|)$.

4) Forwarding Strategy: The enhanced TCC-aware data forwarding strategy is presented in Algorithm 2. Based on the original TCC-aware data forwarding strategy, the enhanced TCC-aware data forwarding strategy adds an extra step that redistributes data copies inside a TCC (Lines $18 \sim 24$). Similar to the original strategy, one extra data copy is created if the node with the highest centrality does not have the data (Lines $18 \sim 22$). Then, all data copies are redistributed in the TCC, to make sure they are carried by a set of nodes with the optimal forwarding capability. Specifically, the command node first computes an optimal node set with the highest set centrality using the method in Section III-B3 (Lines 23). Afterwards, the data copies are forwarded from the original data carries to the nodes in the optimal node set (Lines 24).

IV. Performance Evaluations

In this section, we evaluate the performance of the TCC-aware data forwarding strategies by comparing them with existing data forwarding strategies.

A. Strategies in Comparison

In TCC-aware data forwarding strategies (TCC and Enhanced TCC), the centrality metric is based on the Cumulative Contacting Probability (CCP) [5]. We compare it with the following four forwarding strategies:

- **Compare-and-Forward:** When a node carrying data contacts a node without data, the data carrier forwards the data if the contacted node has a higher forwarding metric (CCP). This strategy has been utilized in [3][2][18].
- **Epidemic** [19]**:** Upon contact, the data carrier forwards data to the contacted node if it does not have the data. This method has the best data delivery ratio but the highest network overhead, which is used as the upper bound.
- **Wait:** The data source does not forward data to any other node until it contacts with the destination. This strategy has the worst delivery ratio but the lowest network overhead, which is used as the lower bound.
- **R3** [20]**:** R3 selects a fixed number of forwarding paths which together achieve the minimum expected delay. For each path, one data copy is created and then forwarded along the path.

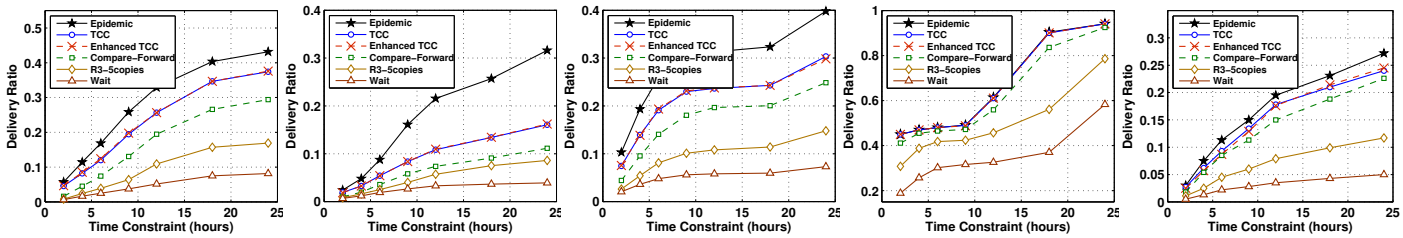
B. Simulation Setup

We evaluate all the data forwarding strategies based on the five traces listed in Table I. For each trace, the first half is used for warm up, based on which we determine the centrality metrics, and the second half is used to evaluate the performance.

For each data item, the source and destination are picked randomly, and the data generation time is randomly chosen in the daytime, since node's activity remains low at night and high at daytime. The experiment is repeated 1000 times for statistical convergence.

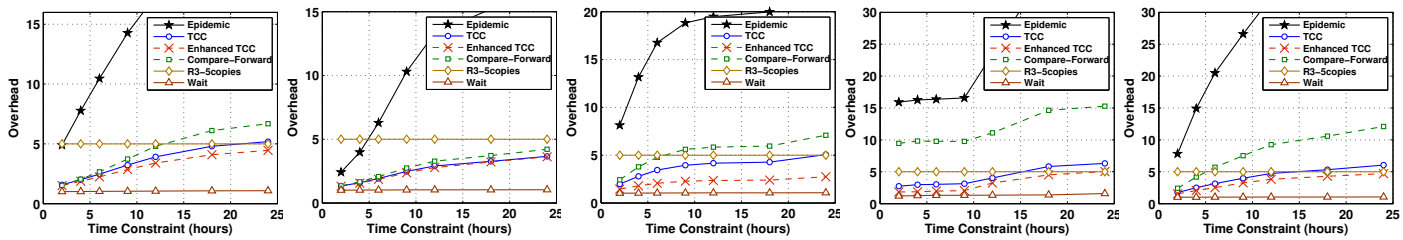
C. Simulation Results

Figure 7 and Figure 8 compare TCC-aware data forwarding strategies (TCC and Enhanced TCC) and other existing strategies. Figure 7 shows data delivery ratio under different time constraints, where delivery ratio is the proportion of data items successfully delivered to the destination before data expire. Figure 8 shows the network overhead under different time constraints, where the overhead is measured by the average number of data copies created in the network for each data item. Generally speaking, TCC-aware data forwarding strategies have the best performance by achieving comparable delivery ratio with Epidemic with much lower network overhead. In the following, we discuss the



(a) Social Evolution (b) Friends & Family (c) Reality Mining (d) Infocom (e) UCSD

Figure 7. Comparisons based on data delivery ratio under different time constraints.



(a) Social Evolution (b) Friends & Family (c) Reality Mining (d) Infocom (e) UCSD

Figure 8. Comparisons based on the number of data copies (overhead) created under different time constraints.

simulation results in detail by first comparing our TCC-aware data forwarding strategies with Compare-and-Forward and R3, and then comparing these two TCC-aware data forwarding strategies.

1) Comparisons with Compare-and-Forward: As shown in Figure 7 and Figure 8, both TCC and Enhanced TCC have better performance than Compare-and-Forward, with higher data delivery ratio and lower network overhead. Specifically, the TCC-aware data forwarding strategies achieve $10\% - 40\%$ higher delivery ratio than Compare-and-Forward in all five traces. This is because by utilizing the contact opportunities inside TCCs, nodes have higher probabilities to forward data to the destination. Moreover, the TCC-aware data forwarding strategies consume $15\% - 50\%$ less network overhead than Compare-and-Forward. The decrease in network overhead is because there is at most one data copy created when a new TCC is formed. However, in Compare-and-Forward, data copies can be created upon every contact. These results demonstrate the effectiveness of TCC-aware data forwarding strategies when compared with strategies designed for opportunistic networks.

2) Comparisons with R3: To compare with R3 [20], we set the number of data copies (paths) to be 5 (R3-5copies) in R3. As shown in Figure 7, the data delivery ratio for R3 with 5 copies is about 50% less than that of TCC and Enhanced TCC. We also test R3 with 10 copies and 20 copies, and find that the delivery ratio does not increase much as the number of data copies increases. This observation is consistent with the results in [20], which also found that the performance does not increase much as the number of data copies is larger. The reason for R3's low delivery ratio is that R3 is based on source routing, and source routing is extremely time-consuming when utilized in networks with opportunistic feature. MSNs with diverse connectivity characteristics have the opportunistic feature, because nodes contact opportunistically outside of TCCs.

3) Comparisons between the TCC-aware data forwarding strategies: From Figure 7 and Figure 8, we can also see that Enhanced TCC consumes less overhead than TCC but achieves similar delivery ratio. This is because, in the enhanced strategy, data copies can be forwarded from the original data carriers to a set of nodes with the highest set centrality; i.e., more effective nodes are selected as data carriers. By choosing a small number of effective nodes to carry data, the enhanced strategy requires less data carriers than the original strategy. Therefore, the number of data copies can be decreased with Enhanced TCC.

V. Related Work

Data forwarding algorithms designed for opportunistic networks are commonly used to forward data in MSNs, and most of them are based on Epidemic [19], where data is flooded upon contacts with other nodes. Later solutions attempt to reduce the number of data copies created by Epidemic, and these strategies are known as controlled flooding [21]. For example, Compare-and-Forward is commonly used to control the data copies created, and the data carrier only forwards data to another node with higher forwarding metric. The forwarding metric measures node's capability of forwarding data to the destination. In some research, the forwarding metrics are determined based on node's contact probability with the destination, such as PROPHET [2] and MaxProp [22]. By further aggregating the networks to social graphs [23], social properties of mobile nodes are analyzed.

Based on this, node centrality [16][18][24] or community based solutions[25][26][27][28] are used as forwarding metrics. However, these strategies are designed for opportunistic networks, and they are not the best solutions for MSNs with diverse connectivity characteristics, since they neglect the multi-hop communication opportunities inside TCCs.

Phe-Neau *et al.* considered the multi-hop communication opportunities around a node's vicinity in [29]. However, in their approach, multi-hop communication opportunities around a node are only utilized when the destination of the data is within the node's vicinity, which is basically a WAIT strategy and then it does not contain mechanisms to make the data reach more nodes and get closer to the destination.

The work by Tie *et al.* [20] considered data forwarding in networks with diverse connectivity characteristics; however, their solution is different from ours and they do not consider the effects of TCCs. They identified packet replication to be the key difference between protocols designed for well-connected networks and sparsely-connected networks, and designed a routing protocol called R3, which determines the number of data copies to be created based on the predicted delays along network paths. R3 is based on source routing; i.e., data copies are forwarded along the pre-determined forwarding path. However, protocols based on source routing are not suitable for networks with opportunistic features, because it is extremely time-consuming to forward data along the pre-determined paths. Moreover, R3 does not consider the effects of TCCs on data forwarding, which is the key contribution of our work.

Other existing algorithms try to modify well-known MANET protocols to make them more adaptive to networks with diverse connectivity characteristics. For example, Raffelsberger *et al.* [30] integrated store-and-forward to MANET protocols. In MANET protocols, a data item is dropped when the routing table does not contain an entry for the destination. With store-and-forward, data is buffered until a route to the destination can be found using MANET protocols. However, their solution lacks a mechanism to choose effective data carriers to deliver data to destination.

There exists some work on analyzing TCCs. For example, [31] demonstrated the size of the giant connected components changes over time. [13][14][15] proved that the property of connected component is closely related with the distribution of node degree. However, they did not examine the detailed structure of TCCs using real traces, and did not consider how to use them to increase the contact opportunities and improve the performance of data forwarding, which is the focus of our work.

VI. Conclusions

In this paper, we designed efficient data forwarding strategies for MSNs with diverse connectivity characteristics, by exploiting the existence of TCCs. We first identified the existence of TCCs and analyzed their properties based on five traces. By treating multi-hop wireless communications inside TCCs as indirect contacts, through theoretical analyses, we showed that the contact opportunities can be significantly increased in all traces. Based on this observation, we designed a TCC-aware data forwarding strategy to improve the performance of data forwarding in MSNs. Then, we enhanced the TCC-aware data forwarding by selecting an optimal set of nodes in the TCC to avoid overlap in their contacts and maximize the data forwarding opportunity with a small number of nodes. Trace-driven simulations showed that our TCC-aware data forwarding strategies outperform existing data forwarding strategies with less network overhead.

References

[1] S. Ioannidis, A. Chaintreau, and L. Massoulié, "Optimal and scalable distribution of content updates over a mobile social network," in *INFOCOM 2009, IEEE*. IEEE, 2009, pp. 1422–1430.

[2] A. Lindgren, A. Doria, and O. Schelén, "Probabilistic routing in intermittently connected networks," *ACM SIGMOBILE CCR*, vol. 7, no. 3, pp. 19–20, 2003.

[3] E. Daly and M. Haahr, "Social network analysis for routing in disconnected delay-tolerant manets," in *Proc. ACM MobiHoc*. ACM, 2007, pp. 32–40.

[4] Y. Zhang and G. Cao, "V-pada: Vehicle-platoon-aware data access in vanets," *Vehicular Technology, IEEE Transactions on*, vol. 60, no. 5, pp. 2326–2339, 2011.

[5] W. Gao, Q. Li, B. Zhao, and G. Cao, "Multicasting in delay tolerant networks: a social network perspective," in *Pro ACM MobiHoc*. ACM, 2009, pp. 299–308.

[6] A. Madan, M. Cebrian, S. Moturu, K. Farrahi, and A. Pentland, "Sensing the health state of a community," *Pervasive Computing*, vol. 11, no. 4, pp. 36–45, 2012.

[7] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland, "Social fmri: Investigating and shaping social mechanisms in the real world," *Pervasive and Mobile Computing*, vol. 7, no. 6, pp. 643–659, 2011.

[8] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, 2006.

[9] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *Mobile Computing, IEEE Transactions on*, vol. 6, no. 6, pp. 606–620, 2007.

[10] M. McNett and G. Voelker, "Access and mobility of wireless pda users," *ACM SIGMOBILE CCR*, vol. 9, no. 2, pp. 40–55, 2005.

[11] Å. Björck, *Numerical methods for least squares problems*. Siam, 1996.

[12] T. Spyropoulos, K. Psounis, and C. Raghavendra, "Spray and wait: an efficient routing scheme for intermittently connected mobile networks," in *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*. ACM, 2005, pp. 252–259.

[13] M. Molloy and B. Reed, "The size of the giant component of a random graph with a given degree sequence," *Combinatorics probability and computing*, vol. 7, no. 3, pp. 295–305, 1998.

[14] W. Aiello, F. Chung, and L. Lu, "A random graph model for power law graphs," *Experimental Mathematics*, vol. 10, no. 1, pp. 53–66, 2001.

[15] M. E. J. Newman, "2 random graphs as models of networks," *Handbook of graphs and networks*, p. 35, 2003.

[16] P. Hui, J. Crowcroft, and E. Yoneki, "Bubble rap: Social-based forwarding in delay-tolerant networks," *IEEE Transactions on Mobile Computing*, vol. 10, no. 11, pp. 1576–1589, 2011.

[17] S. Martello and P. Toth, *Knapsack problems: algorithms and computer implementations*. John Wiley & Sons, Inc., 1990.

[18] W. Gao and G. Cao, "On exploiting transient contact patterns for data forwarding in delay tolerant networks," in *Proc. ICNP*. IEEE, 2010, pp. 193–202.

[19] A. Vahdat, D. Becker *et al.*, "Epidemic routing for partially connected ad hoc networks," Technical Report CS-200006, Duke University, Tech. Rep., 2000.

[20] X. Tie, A. Venkataramani, and A. Balasubramanian, "R3: robust replication routing in wireless networks with diverse connectivity characteristics," in *Mobicom*. ACM, 2011, pp. 181–192.

[21] K. A. Harras, K. C. Almeroth, and E. M. Belding-Royer, "Delay tolerant mobile networks (dtmns): Controlled flooding in sparse mobile networks," in *NETWORKING*. Springer, 2005, pp. 1180–1192.

[22] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine, "Maxprop: Routing for vehicle-based disruption-tolerant networks." in *INFOCOM*, vol. 6, 2006, pp. 1–11.

[23] T. Hossmann, T. Spyropoulos, and F. Legendre, "Know thy neighbor: Towards optimal mapping of contacts to social graphs for dtn routing," in *INFOCOM, 2010 Proceedings IEEE*. IEEE, 2010, pp. 1–9.

[24] W. Gao and G. Cao, "User-centric data dissemination in disruption tolerant networks," in *INFOCOM, 2011 Proceedings IEEE*, 2011, pp. 3119–3127.

[25] P. Hui and J. Crowcroft, "How small labels create big improvements," in *PerCom Workshops*. IEEE, 2007, pp. 65–70.

[26] N. Nguyen, T. Dinh, S. Tokala, and M. Thai, "Overlapping communities in dynamic networks: their detection and mobile applications," in *Proc. ACM MobiCom*. ACM, 2011, pp. 85–96.

[27] Z. Lu, Y. Wen, and G. Cao, "Community detection in weighted networks: Algorithms and applications," in *PerCom 2013, IEEE*, 2013, pp. 179–184.

[28] X. Zhang and G. Cao, "Transient community detection and its application to data forwarding in delay tolerant networks," in *Proc. of IEEE ICNP*, 2013.

[29] T. Phe-Neau, M. D. De Amorim, V. Conan *et al.*, "The strength of vicinity annexation in opportunistic networking," in *Fifth International Workshop on Network Science for Communication Networks (NetSciCom 2013)*, 2013.

[30] C. Raffelsberger and H. Hellwagner, "A hybrid manet-dtn routing scheme for emergency response scenarios," in *PerCom Workshops*. IEEE, 2013.

[31] A. Pietilänen and C. Diot, "Dissemination in opportunistic social networks: the role of temporal communities," in *Proc. ACM MobiHoc*. ACM, 2012, pp. 165–174.

APPENDIX A. PROOF OF THEOREM 1

Proof: Based on the condition that the distributions of node degree and TCC size follow exponential distributions, we compute the number of direct contacts and TCC contacts.

The number of direct contacts, denoted as m_d, can be determined by the network size N and the distribution of node degree with exponential constant k^*. Let $p(k)$ denote the probability mass function (PMF) of degree K. $p(k)$ can be computed from the CCDF $P(K > k)$ as

$$\begin{aligned} p(k) &= P(K > k-1) - P(K > k) \\ &= e^{-\frac{k-1}{k^*}} - e^{-\frac{k}{k^*}} = e^{-\frac{k-1}{k^*}}(1 - e^{-\frac{1}{k^*}}). \end{aligned}$$

for $k \geq 1$. Then, m_d can be computed as half of the number of total degrees:

$$m_d = \frac{1}{2}\sum_{k=1}^{\infty}(N \cdot p(k) \cdot k) = \frac{N}{2(1 - e^{-\frac{1}{k^*}})}. \tag{11}$$

We next compute the number of TCC-contacts, m_t, with the network size N and the distribution of TCC size which has exponential constant s^*. The PMF $p(s)$ of TCC size S can be calculated from CCDF $P(S > s)$ as

$$\begin{aligned} p(s) &= P(K > s-1) - P(K > s) \\ &= \begin{cases} e^{-\frac{s-1}{s^*}}(1 - e^{-\frac{1}{s^*}}) & s \geq 3 \\ 1 - e^{-\frac{2}{s^*}} & s = 2. \end{cases} \end{aligned}$$

We assume the total number of TCCs inside the network is N_{TCC}. With the distribution of TCC sizes $p(s)$, the number of TCC with size s is $N_{TCC} * p(s)$. The number of TCC-contacts inside a TCC with size s is $\binom{s}{2} = \frac{s(s-1)}{2}$. The total number of TCC-contacts inside the network is:

$$\begin{aligned} m_t &= \sum_{s=2}^{\infty}(N_{TCC} \cdot p(s) \cdot \frac{s(s-1)}{2}) \\ &= N_{TCC} \cdot \frac{E(S^2) - E(S)}{2} \end{aligned} \tag{12}$$

The value of N_{TCC} can be determined by the network size N. With the distribution of TCC size $p(s)$, the number of TCCs with size s is $N_{TCC} \cdot p(s)$. Since a node belongs to one TCC, the summation of all TCC sizes is the network size N. Thus, we have

$$N = \sum_{s=2}^{\infty}(s \cdot N_{TCC} \cdot p(s)) = N_{TCC} \cdot E(S).$$

Therefore, $N_{TCC} = \frac{N}{E(S)}$. With

$$\begin{aligned} E(S) &= 2(1 - e^{-\frac{2}{s^*}}) + \sum_{s=3}^{\infty}(s \cdot e^{-\frac{s-1}{s^*}}(1 - e^{-\frac{1}{s^*}})) \\ &= \frac{1}{1 - e^{-\frac{1}{s^*}}} + 1 - e^{-\frac{1}{s^*}}, \\ E(S^2) &= 4(1 - e^{-\frac{2}{s^*}}) + \sum_{s=3}^{\infty}(s^2 \cdot e^{-\frac{s-1}{s^*}}(1 - e^{-\frac{1}{s^*}})) \\ &= \frac{1 + e^{-\frac{1}{s^*}}}{(1 - e^{-\frac{1}{s^*}})^2} + 3 - 3e^{-\frac{1}{s^*}}, \end{aligned}$$

Equation (12) becomes

$$\begin{aligned} m_t &= \frac{N}{E(S)} \cdot \frac{E(S^2) - E(S)}{2} = N \cdot \frac{E(S^2) - E(S)}{2E(S)} \\ &= N \cdot \frac{e^{-\frac{1}{s^*}} + (1 - e^{-\frac{1}{s^*}})^3}{1 - e^{-\frac{1}{s^*}} + (1 - e^{-\frac{1}{s^*}})^3}. \end{aligned} \tag{13}$$

The ratio of TCC-contacts to direct contacts is

$$\frac{m_t}{m_d} = 2(1 - e^{-\frac{1}{k^*}}) \cdot \frac{e^{-\frac{1}{s^*}} + (1 - e^{-\frac{1}{s^*}})^3}{1 - e^{-\frac{1}{s^*}} + (1 - e^{-\frac{1}{s^*}})^3}. \tag{14}$$

which is determined only by the two exponential constants k^* and s^*, independent of the network size N. With $\frac{m_t}{m_d} > 1$, we have

$$k^* < \frac{1}{\log \frac{2 \cdot (e^{-\frac{1}{s^*}} + (1 - e^{-\frac{1}{s^*}})^3)}{3e^{-\frac{1}{s^*}} - 1 + (1 - e^{-\frac{1}{s^*}})^3}} \tag{15}$$

If $\hat{k}^* = 1/\log \frac{2 \cdot (e^{-\frac{1}{s^*}} + (1 - e^{-\frac{1}{s^*}})^3)}{3e^{-\frac{1}{s^*}} - 1 + (1 - e^{-\frac{1}{s^*}})^3}$, as long as $k^* < \hat{k}^*$, the number of TCC-contacts is more than the direct contacts, which increases the contact opportunities. ■