

节点相似度感知的 DTN 交叠社区结构检测机制

王 峥^{1,2}, 王于波^{1,2}, 朱承治³, 任孝武^{1,2}, 庞振江^{1,2}

(1 北京智芯微电子科技有限公司, 北京 100192; 2 国网信息通信产业集团有限公司,
北京 102211; 3 国网浙江省电力公司, 浙江 杭州 310007)

摘 要: 提出了一种基于节点相似度的交叠社区检测算法, 利用相同社区内节点较高的相似度特性和不同节点间相似度来初步确定节点间的模糊关系, 在此基础上计算节点的熵值以确定节点的社区归属. 数值结果表明所提出的检测机制与经典的社区检测机制相比较, 能显著提高检测的准确性.

关键词: DTN; 交叠社区检测; 节点相似度; 熵值

中图分类号: TP393. 04

文献标识码: A

文章编号: 1000-7180(2017)12-0074-05

Similarity-awareness Based Overlapping Community Detection Mechanism in Delay Tolerant Network

WANG Zheng^{1,2}, WANG Yu-bo^{1,2}, ZHU Cheng-zhi³,
REN Xiao-wu^{1,2}, PANG Zhen-jiang^{1,2}

(1 Beijing Smartchip Microelectronics Technology Co., Ltd., Beijing 100192, China;
2 State Grid Information & Telecommunication Co., Ltd., Beijing 102211, China;
3 State Grid Zhejiang Electric Power Corporation, Hangzhou 310007, China)

Abstract: This paper proposes a new algorithm based on the node similarity to detect the overlapping community. Firstly, the fuzzy relation between nodes is initially determined by using the similarity degree between nodes. Then, by computing node entropy, the communities which nodes belong to can be determined. The numerical results show that the proposed detection mechanism compared with classic community detection mechanism can significantly improve the precision of detection.

Key words: delay tolerant networks; overlapping community detection; node similarity; entropy

1 引言

不同于传统的移动自组织网络, 延迟容忍网络(delay tolerant networks, DTN)^[1-3]中源节点和目的节点间无需预先建立完整的端到端路径, 利用节点运动过程带来的相遇机会, 以更加灵活的“存储-携带-转发”方式实现节点间数据转发. 为了节省自身的资源, 节点只转发与自身相关性强的数据. 同时, 大量实测数据^[4-5]分析表明, DTN 中节点的运动具有一定规律性, 呈现出极强的社会特征, 即“小世界, 大世界”特性^[6-7]. 此种模式下, 网络节点按照一定的规则以自组织的方式形成多个社区, 但每个社

区边界段将以一定的概率呈现出模糊、重叠现象, 即少数节点同时存在于多个社区中, 从而形成相互重叠的社区(Overlapping Communities)^[8].

目前采用的社区检测算法主要分为在线式社区检测算法和离线式社区检测算法^[9]. 传统的有线网络拓扑结构固定不变, 节点不受能量和缓存的制约, 且每个节点均可记录整个网络的拓扑信息, 故采用离线式社区检测机制. 不同于前者, 由于 DTN 具有间断连接、拓扑结构高度动态和节点通信覆盖范围小等特性, 将导致节点无法实时有效地获取相关属性信息, 且无法共享全局信息, 即离线式社区检测算法在 DTN 中适应性较差.

针对上述存在的问题,文献[10]提出一种基于节点分裂的算法,将重叠社区检测转化为非重叠社区检测,进而对社区进行划分.该过程中,节点是否分裂取决于分裂系数,带来了较高的时间复杂度,且全局信息的获取难以实现,故该文社区检测的准确性降低,不适合处理大型现实的 DTN.文献[11]提出一种基于贪婪算法原理的算法,采用试探优化方法将网络中的节点随机划分为多个已知规模的社区.但该算法要求已知网络社区的确切规模才能得以应用.然而 DTN 中节点的相关信息具有高度不完备性,因此网络社区的规模难以准确度量.文献[12]通过周期性选取连接度最大的节点为社区中心节点,将其通信范围内的节点自动归属为该中心节点的社区成员,以此方法蔓延至全网络节点.但选取合适的周期难度较大,且网络中部分节点不能对网络连通性起到关键支撑.因此该算法缺乏对网络中心节点的解释能力,易使检测算法在中心节点的选取上产生误判,从而导致社区划分产生偏差.文献[13]分析节点在运行的过程中记录与其他节点的相遇情况及节点间的连接紧密程度来判断节点自身所属社区.但该机制采用离线式检测算法,易使已完成划分的社区结构难以为继且该算法需要人为预先设定社会图中边的权重,主观性较强,使得精度降低.

针对 DTN 中重叠社区难以检测的问题,本文在分析现有社区检测算法的基础上,考虑网络信息的不完备性及拓扑结构的高动态性,从社会性网络特征及使用概念角度出发,提出了一种面向 DTN 的重叠社区检测机制.

2 节点相似度估计方法

为了确定邻居列表中的节点元素,如图 1 所示,节点 u 与其连接的节点分别是 $u_1, u_2, \dots, u_n$, 且将历史信息保存在节点 u 的连接列表 $Ln(u)$.

节点 u_1	节点 u_2	...	节点 u_n
(u_1, c_{u_1})	(u_2, c_{u_2})	...	(u_n, c_{u_n})

图 1 节点连接列表

其中, (u_n, c_{u_n}) 分别表示与节点 u 连接的节点 ID 号和相应的连接次数,且 $c_{u_n} = ET_{Ln(u)}$.

为了准确估计节点间相似度的同时降低算法复杂度,本文通过分离那些连接紧密度较高的节点,以此建立节点的邻居列表,从而确定邻居列表中的节点元素.邻居节点用式(1)表示为

$$\overline{ET_{Ln(u)}} = \sum_1^n ET_{Ln_x(u)} / n \quad (1)$$

式中, n 表示节点 u 连接列表中节点个数, $ET_{Ln_x(u)}$ 表示节点 u 的连接列表中第 x 个节点的连接次数.若 $ET_{Ln_x(u)} > \overline{ET_{Ln(u)}}$ 则将该节点 ID 号及对应的连接次数存放在节点 u 的邻居列表 $Lm(u)$ 中,以此来获得节点 u 的邻居节点.

通过上述方法获得节点 i 的邻居列表 $Ls(i)$ 后,列表中所存储的 s 个邻居节点与本节点连接紧密度都较高,但它们对节点 i 的影响程度有所不同.为了明显区分邻居节点对节点 i 的影响程度,对邻居列表中每个节点赋予一个相应的权重值 W ,来准确量化其影响程度.节点 i 的邻居列表可用一个一维向量表示为 $D_{is} = D(N_1, R_{i1}; N_2, R_{i2}; \dots; N_s, R_{is})$, 其中 $N_1, N_2, \dots, N_s$ 表示节点 i 的邻居节点, $R_{i1}, R_{i2}, \dots, R_{is}$ 对应节点的连接次数. W_k 表示邻居节点 $N_k (1 \leq k \leq s)$ 相对于节点 i 的权重.根据节点 N_k 连接次数占节点 i 总连接次数的比重,以获得节点 N_k 相对于节点 i 的权重 W_k ,如式(2)所示.

$$W_k = R_k / \sum_{i=1}^s R_k \quad (2)$$

通过上述可获得节点 i 每个邻居节点相对于它的影响程度为 $d_{is} = D(W_{i1}, W_{i2}, \dots, W_{is})$. 其中 s 表示节点 i 中邻居节点的个数.

若两节点间有一个共同连接节点,且它们与该节点的连接次数相近,则表明两个节点间运动轨迹和到访同一社区的频率皆相近^[14-15].因此,在估计节点间相似度时,可采用向量模的计算方法,将节点邻居列表中各个元素看作单一向量,逐一将两个节点邻居列表中每个节点的权重值 W 进行向量相乘.从而,通过 d_{is} 与 d_{jt} 向量乘积可得到节点 i 与节点 j 间的共同邻居节点对它们的影响程度,如式(3)所示.

$$d_{is}^T d_{jt} = \sum_{k=1}^{s+t} W_{ik} \times W_{jk} \quad (3)$$

为了准确估计节点间的相似度,需综合考虑节点每个邻居节点对其自身的影响程度,用式(4)和式(5)表示如下.

$$|d_{is}| = \sqrt{\sum_{k=1}^s W_{ik}^2} \quad (4)$$

$$|d_{jt}| = \sqrt{\sum_{k=1}^t W_{jt}^2} \quad (5)$$

综上所述,由式(2) ~ (5) 可计算出节点 i 和节点 j 间的相似度 $SV(\text{Similar Value})$,如式(6)所示.

$$SV(i, j) = (d_{is} \times d_{jt}) / (|d_{is}| \times |d_{jt}|) \quad (6)$$

式中, $|d_{is}| \times |d_{jt}|$ 表示节点 i 和节点 j 的共同邻居节点的个数.

3 节点重要程度感知方法

DTN 中节点具有显著的异构性,且处于不同地位的节点对网络性能的影响程度不同.因此,为了能够准确地划分社区结构,本文综合考虑节点间的相似度和每个节点所处社会地位的重要程度(Node Importance, NI).此外,如果节点 i 处于“要塞”位置,则网络中较多的路径都要经过节点 i ,则节点 i 所携带的消息成功转发给目的节点的概率越大.因此,综合考虑感知节点的网络连通能力(Connectivity Capability, CC)和位置重要性(Position Importance, PI),有效地预测了节点在网络中的重要程度.

3.1 节点连通能力估计

在相同时间内,若连接次数相同,则连接规律性越强的两节点间具有更紧密的联系.因此,在考虑连接次数的同时,还需对与其他节点相遇的稳定性进行度量.节点 i 连通能力 CC_i 可用式(7)表示.

$$CC_i = T / \sum_{x=0}^M f(x)^2 \quad (7)$$

式中, T 为以小时为单位的网络运行时间. $f(x) = t_{\text{next}} - t_{\text{last}}$, 为相遇节点的相遇间隔时间; t_{last} 为节点 i 遇到前一节点的时间; t_{next} 为遇到下一节点的时间; $\sum_{x=0}^M f(x)^2$ 随间隔时间的减小而减小.因此,当节点 i 与其他节点相遇时间间隔越短,且对应的相遇时间间隔方差越小时, $\sum_{x=0}^M f(x)^2$ 的累加速度越缓慢,则节点 i 在网络中具有较强的连通能力.

3.2 节点位置重要性估计

为了衡量节点 i 在网络中所处位置的重要性 PI_i ,本文采用在单位时间内节点 i 所遇到节点的个数来进行衡量,如式(8)所示.

$$PI_i = \sum_{j \neq i, j \in G} C_{ij} / T \quad (8)$$

式中, G 表示为该网络; C_{ij} 表示两节点间的连接次数.

3.3 节点重要程度估计

通过以上分析,综合考虑节点在网络中的连通能力和位置重要性来评估一个节点的重要程度.因此,根据式(7)和式(8),本文将节点 i 的重要程度 NI_i 的量化方法定义如式(9)所示.

$$NI_i = PI_i \times CC_i \quad (9)$$

显然,节点 NI 值较高,表明该节点所处网络位

置越重要,从而其社会地位就越高,同时该节点在网络中的连通能力则较大,进而使气对消息的传输有较大的贡献.

4 重叠社区检测算法

本文提出的重叠社区检测算法 EB 包括两个阶段:(1)通过感知节点间的相似度构建网络中模糊社区结构,同时根据节点的重要程度,进一步对该模糊社区中的节点进行准确定位,形成初步社区框架;(2)分析节点关于每个模糊社区的熵值,从而检测出模糊社区中重叠的部分.实际网络中,节点间的重要程度存在差异,导致网络社区结构不能准确地建立.故在通过节点的相似性构建模糊社区后,还需发掘出每个模糊社区中重要程度最高的节点,以确定节点的所属社区,形成初步社区结构.由于真实的 DTN 中部分社区间常产生重叠现象,一些节点同时属于多个社区,因此,在第二阶段中,重点衡量出该类节点的社区归属程度,以此确定节点的准确归属,检测出网络的重叠社区结构.

根据熵值理论可知,假设 $G' = (V', E')$ 是全网网络 $G = (V, E)$ 的一个社区,则将 G 中的任意节点 v 与网络中其他节点的连接边 (u, v) 可以分为两类,即与社区 G' 内节点的连接边 $\{(u, v) \mid u \in V', (u, v) \in E'\}$ 和与社区 G' 外节点的连接边 $\{(u, v) \mid u \notin V', (u, v) \in E'\}$.节点 v 与社区 G' 内节点的连接边 $\{(u, v) \mid u \in V', (u, v) \in E'\}$ 的总数可用式(10)表示为

$$N_{G'}(v) = \sum_{u \in G', u \neq v} n_{uv} \quad (10)$$

式中, n_{uv} 为节点 v 与节点 u 间的连接边数.同理可知,节点 v 在整个网络中与其他节点的连接边数为 $N_G(v) = \sum_{u \in G, u \neq v} n_{uv}$.由 $N_{G'}(v)$ 与 $N_G(v)$ 的比值可推出节点 v 属于社区 G' 的概率 $P(v \in V')$,如式(11)所示.

$$P(v \in V') = \frac{|N_{G'}(v)|}{|N_G(v)|} \quad (11)$$

若节点 v 与不在社区 G' 内的网络中其他节点的连接边数总和为 $N_{G \setminus G'}(v) = \sum_{u \in G, u \notin G', u \neq v} n_{uv}$,同理可推出节点 v 不属于社区 G' 的概率 $P(v \notin V')$,如式(12)所示.

$$P(v \notin V') = \frac{|N_{G \setminus G'}(v)|}{|N_G(v)|} \quad (12)$$

根据上述推导出的 $P(v \in V')$ 与 $P(v \notin V')$ 可计算出节点 v 对于社区 G' 的熵值,如式(13)所示.

$$e_v(G') = -p_i(v) \log_2 p_i(v) - p_o(v) \log_2 p_o(v) \quad (13)$$

根据熵值能够有效地分析出节点对于给定社区的归属程度. 本文对归属程度设置一个阈值 γ , 若节点 v 对于社区 G' 的熵值 $e_v(G')$ 小于阈值 γ , 则说明此时节点 v 与社区 G' 中节点的连接稳定性较高, 不容易脱离该社区, 从而认定该节点应归属于社区 G' , 进而检测出重叠社区结构.

5 数值分析

5.1 仿真环境设置

本算法是在 ONE(Opportunistic Network Environment)仿真平台^[14]验证的. 将本文算法与 K-CLIQUE、HCDA 和 WNA 进行比较以突出本文算法的有效性和可行性. 由理论分析与实测数据分析可知, 真实网络中独立节点是网络组成的主体成员, 但网络中存有少于 30% 的具有交叠身份的节点. 故本文在放入了少量交叠节点的混杂化网络环境下进行社区检测, 这些交叠节点以相近的概率往返于两个社区, 更加贴近实际网络.

5.2 度量方法

本文采用一种更客观的度量方法: 归一化互信息对社区检测结果进行分析. 定义一个混淆矩阵 $\mathbf{N}$, 在该矩阵中, 每一行对应一个“真实”社区, 每一列对应一个所使用的算法来“挖掘出”的社区. $\mathbf{N}$ 的元素 N_{ij} 表示“真实”社区 i 和发现社区 j 重合的顶点个数. 矩阵第 i 行的元素之和记为 $N_{i\cdot}$, 矩阵第 j 列的元素之和记为 $N_{\cdot j}$, $\mathbf{N}$ 的所有元素之和记为 S . 用 C_R 表示“真实”社区的个数, 用 C_F 表示发现的社区个数. 基于信息论, 得到发现的社区划分 F 和真实的社区划分 R 之间相似度的一种度量:

$$NMI(R, F) = \frac{-2 \sum_{i=1}^{C_R} \sum_{j=1}^{C_F} N_{ij} \log(\frac{N_{ij} S}{N_{i\cdot} N_{\cdot j}})}{\sum_{i=1}^{C_R} N_{i\cdot} \log(\frac{N_{i\cdot}}{S}) + \sum_{j=1}^{C_F} N_{\cdot j} \log(\frac{N_{\cdot j}}{S})} \quad (14)$$

当发现的划分和网络的真实划分一致时, NMI 取值为 1. 如果发现的划分和网络的真实划分完全独立时, NMI 取值为 0. 本文将使用 NMI 评测方法来度量 EB 的好坏.

5.3 仿真数值分析

图 2 描述了四种检测算法随网络运行时间的变化. 可以看出, 随着网络的运行, 节点逐渐掌握了网络中的有效信息, 故四种检测算法的准确率都逐渐

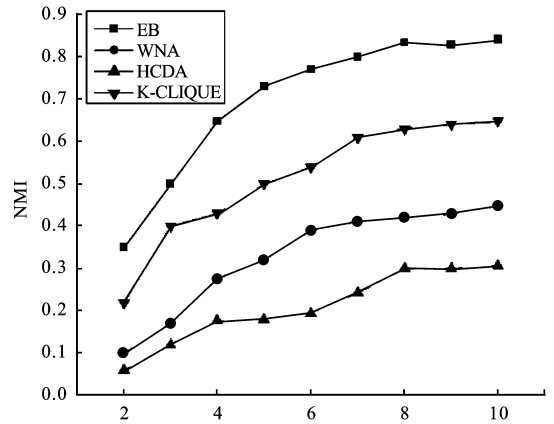


图 2 NMI 与运行时间的关系

上升, 从图中可以看出 EB 算法 NMI 值高于其他三种算法, 这是因为 EB 算法在确定模糊关系时只需进行线性比较, 有利于式(5)的关系确定, 此外, 节点在网络运行的同时携带了其感知到的本社区的最高重要程度节点, 极大地加快了其他节点与最高重要程度节点的相似度计算, 从而缩短了其认知到自身所属社区的时间. 由于四种检测算法对节点相遇顺序具有敏感性, 部分节点在初期遇到不同社区的节点, 导致误判归属到了其他社区, 故随着网络运行的深入, 四种算法的检测准确率上升变慢, 最后趋于平缓. K-CLIQUE 和 WNA 均为集中式算法, 因其分享网络相关信息的延迟, 而导致检测效率低于 EB 算法. 且这两种算法对社区检测仅限于对邻居节点的判断, 导致其准确率低. 从图中可以看出当网络运行到 10 小时, EB 算法的 NMI 值高于 WNA 算法 39.3%, 高于 K-CLIQUE 算法 21%.

图 3 描述了 NMI 随社区内节点访问社区外节点概率变化的情况. 从图中可以看出, 由于节点访问

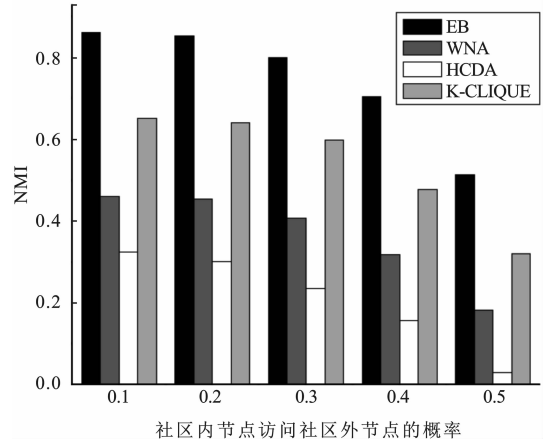


图 3 NMI 与社区内节点访问社区外节点的概率关系

外社区节点的概率增大,各个社区间的独立性与完整性逐渐被模糊化,四种算法的 NMI 值都随着节点访问社区外节点概率的增大而降低. 但 EB 检测算法仍然高于其他算法. 主要由于 EB 检测算法在使用节点连接列表计算与其他节点相似度时,采取了式(1)的计算方法,减少了其他社区节点的相关信息被记录在邻居节点列表的机会,降低了其他社区节点对 EB 算法的干扰. 而其他三种算法均是使用所有的与其他节点连接信息来计算社区的归属,受到外社区节点连接情况相关信息的干扰明显增加,降低了检测的准确性.

6 结束语

为使社区检测算法合理化,使用熵值分析各个节点的社区归属情况. 利用每个社区中节点的中心程度存在差异及同一社区节点具有较高相似度这两个特性,来确定节点间的模糊关系,再通过计算节点的熵值来确定节点的社区归属. 经仿真验证,此 EB 算法能更加高精度地鉴别出每个节点的社区归属.

参考文献:

- [1] Dapeng W, Hongpei Z, Honggang W, et al. Quality of protection (QoP)-driven data forwarding for intermittently connected wireless networks[J]. IEEE Wireless Communication, 2015, 22(4): 66-73.
- [2] Feng M, Mao S, Jiang T. Joint duplex mode selection, channel allocation, and power control for full-duplex cognitive femtocell networks[J]. Digital Communications and Networks, 2015, 1(1): 30-44.
- [3] Soares V N G J, Rodrigues J J P C, Farahmand F. GeoSpray: A geographic routing protocol for vehicular delay-tolerant networks [J]. Information Fusion, 2014, 15: 102-113.
- [4] Pan H, Crowcroft J, Yoneki E. BUBBLE Rap: Social-based forwarding in delay-tolerant networks[J]. IEEE Transactions on Mobile Computing, 2011, 10(11): 1576-1589.
- [5] Pan H, Chaintreau A, Scott J, et al. Pocket switched networks and human mobility in conference environments[C]//Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking. New York, USA: ACM, 2005: 244-251.
- [6] Wu Dapeng, Yanyan Wang, Wang Honggang, et al. Dynamic coding control in social intermittent connectivity wireless networks[J]. IEEE Transactions on Vehicular Technology, 2016, 65(9): 7634-7646.

- [7] Zhu Y, Xu B, Shi X, et al. A survey of social-based routing in delay tolerant networks: positive and negative social effects[J]. IEEE Communications Surveys & Tutorials, 2013, 15(1): 387-401.
- [8] Wang R, Yang H, Wang H, et al. Social overlapping community aware neighbor discovery for D2D communications[J]. IEEE Wireless Communications, 2016, 23(4): 28-34.
- [9] Qi G J, Aggarwal C C, Huang T S. Online community detection in social sensing [C]//Proceedings of the sixth ACM international conference on Web search and data mining. New York, USA: ACM, 2013: 617-626.
- [10] Gregory S. Fuzzy overlapping communities in networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2010, 2(2): P02017.
- [11] Karapetyan D, Gutin G. Lin-Kernighan heuristic adaptations for the generalized traveling salesman problem [J]. European Journal of Operational Research, 2011, 208(3): 221-232.
- [12] Yang W D. Weight-based clustering algorithm for mobile ad hoc network[C]//2011 Cross Strait Quad-Regional Radio Science and Wireless Technology Conference (CSQRWC). Harbin, China: IEEE, 2011, 1: 787-791.
- [13] Yan B, Gregory S. Detecting community structure in networks using edge prediction methods[J]. Journal of Statistical Mechanics: Theory and Experiment, 2012, 2012(9): P09008.
- [14] 吴大鹏, 孔晓龙, 王汝言, 等. 节点重要程度感知的分布式 DTN 结构检测机制[J]. 计算机应用研究, 2012, 29(7): 2645-2649.
- [15] Jolobe O M P. Pharmacogenetic profiling might meet the challenge[J]. Journal of internal medicine, 2011, 270(5): 494-495.

作者简介:

王 峰 男, (1983-), 博士, 工程师. 研究方向为 RFID、传感器技术以及在智能电网等行业领域的应用.

E-mail: wangzhengzhixin123@163.com.

王于波 男, (1969-), 硕士, 高级工程师. 研究方向为信号处理与专用集成电路设计.

朱承治 男, (1977-), 博士, 高级工程师. 研究方向为电力系统及其自动化.

任孝武 男, (1983-), 工程师. 研究方向为电力物联网产品研究及应用.

庞振江 男, (1978-), 硕士, 工程师. 研究方向为数据融合与信息安全.