第43卷 第6A期
2016年6月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

基于频繁闭图关联规则的AS级Internet链路预测方法

张岩庆 陆余良 杨国正
(电子工程学院网络工程系 合肥 230037)

摘 要 目前大多数链路预测方法都是针对丢失链路的结构性预测,缺乏针对未来时刻网络链路的时序性预测,为此提出了一种基于频繁闭图关联规则的链路预测方法。将形式化后的动态网络划分为训练集和测试集,基于Apriori思想从训练集中提取频繁闭图,并根据频繁闭图的时间间隔建立时延分布矩阵,用于表征频繁闭图之间的时序关联规则,在此基础上预测测试集中的网络结构。将该方法运用于不同时间尺度下的AS级Internet动态网络中,结果表明,该方法能够以很高的精确率预测波动型动态网络的链路。

关键词 链路预测,频繁闭图,时序关联,AS级Internet,动态网络
中图法分类号 TP393.4 **文献标识码** A

Link Prediction of AS Level Internet Based on Association Rule of Frequent Closed Graphs

ZHANG Yan-qing LU Yu-liang YANG Guo-zheng
(Department of Network Engineering, Electronic Engineering Institute, Hefei 230037, China)

Abstract The existing link prediction methods are mostly focused on structure link prediction like missing links, but few are about temporal link prediction according to unknown links in future, therefore a link prediction method based on association rules of frequent closed graphs was proposed. Dynamic networks are divided into training set and test test, and frequent closed subgraphs are extracted from training set based on Apriori algorithm, thus time-lag distribution matrix is built to represent the temporal association rules between frequent closed graphs, and then the structure in test set is predicted. The link prediction method was used in the dynamic networks of AS level Internet at different time scales, and experimental results show that this method can efficiently predict links in wavery dynamic networks with high precision.

Keywords Link prediction, Frequent closed graph, Temporal association, AS level Internet, Dynamic networks

1 引言

互联网可以抽象成一个由成千上万个自治域(Autonomous System, AS)相互连接构成的复杂网络,称为AS级Internet。AS之间的连接关系随着时间不断改变,使得AS级Internet具有动态网络的演化特征。AS级Internet的链路预测是指通过已知时间内的网络结构预测未知时刻的网络结构,是网络科学领域的一个重要研究方向,相关研究具有重大的实际应用价值,比如进行网络态势预警、评估网络异常事件效果、揭示网络级联失效机理、分析网络节点关系[1]等。

传统的链路预测方法通常只考虑节点属性信息[2,3],这种方法的局限性在于可靠数据的获取难度大,无法确保节点信息的真实性。相对于节点属性信息,网络拓扑结构信息的获取更容易,而且通用性更强,因此基于网络结构信息的链路预测方法逐渐成为研究的重点。Liben-Nowell等人[4]提出了一种基于网络节点相似性的链路预测方法,验证了拓扑结构信息在链路预测中的有效性;Lahiri等人[5]提出了一种自适应的流式算法,能够以较高的准确率在很小的时间粒度中预测未来时刻的网络结构;Clauset等人[6]在《自然》上发表了一篇利用网络分层结构进行链路预测的方法,其运用最大似然估计的方法推断网络的分层结构,并且在具有显著分层结构的网络中取得了理想的预测效果。

上述链路预测算法主要存在以下两点不足:

1)算法的复杂度高,无法处理大规模的网络集合;

2)算法无法提取网络结构内在的时序关联规则。

在已有研究的基础上,本文提出了一种基于频繁闭图关联规则的链路预测算法,将动态网络中的每个网络快照表示为一系列频繁闭图[7],通过挖掘频繁闭图之间的时序关联规则预测未知时刻的网络结构。频繁闭图完整保存了动态网络的有效结构信息,在实际应用中具有很高的效能比,而且将链路预测的研究对象由边替换为频繁闭图,能够极大地降低计算量。

2 链路预测问题的形式化描述

2.1 基本定义

定义1(动态网络) 动态网络是时间上的有序图集,记

本文受国家自然科学基金(61405248,61503394),安徽省青年科学基金(1408085QF131,1508085QF121)资助。
张岩庆(1988—),男,博士生,主要研究方向为网络态势感知、智能信息处理,E-mail:benqer@126.com;陆余良(1964—),男,教授,博士生导师,主要研究方向为计算机网络安全技术。

为 $G=(V,E,T)=\langle G_1,G_2,\cdots,G_T\rangle=\{G_t\}$。其中，$T$ 是动态网络的快照个数，$V=\{V_t\mid t\in T\}$ 是所有时刻节点集的并集，$E=\{E_t\mid t\in T\}$ 是所有时刻边集的并集，$G_t=(V_t,E_t)$ 是 t 时刻的网络快照，V_t 是 t 时刻的节点集，V_t 中的所有节点是唯一的，E_t 是 t 时刻的边集，边数 $|E_t|$ 表示 t 时刻网络快照的大小。

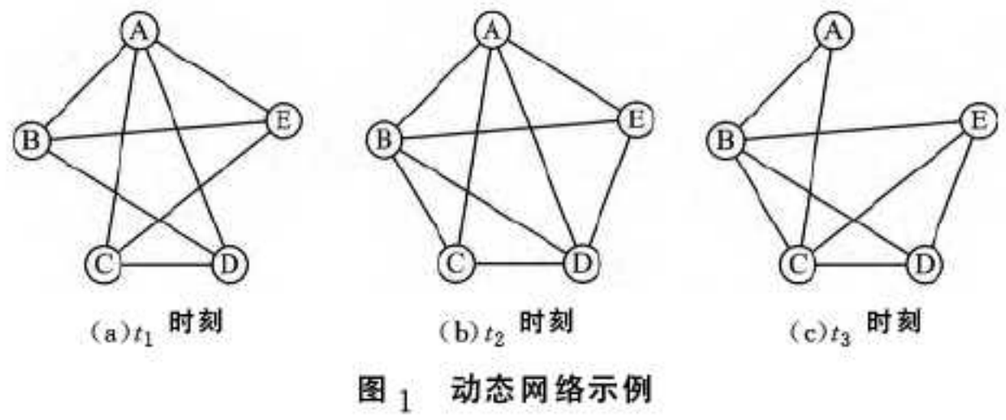
定义 2（子图） 给定两个图 $G_1=(V_1,E_1)$ 和 $G_2=(V_2,E_2)$，如果 $V_2\subseteq V_1$ 且 $E_2\subseteq E_1$，则 G_2 称为 G_1 的子图，记为 $G_2\subseteq G_1$。

定义 3（支持度） 给定动态网络 $G=\langle G_1,G_2,\cdots,G_T\rangle$，则任意图 G' 的支持度 $support(G',G)=\dfrac{|\{t|G'\subseteq G\}|}{T}$。

定义 4（频繁子图） 给定动态网络 $G=\langle G_1,G_2,\cdots,G_T\rangle$ 和任意图 G'，如果 $support(G',G)\geqslant minsup$，则 G' 是 G 的一个频繁子图。其中，$0<minsup\leqslant 1$ 是预设的最小支持度门限值。

定义 5（频繁闭图） 给定动态网络 G 的频繁子图集合 F 和一个子图 $f\in F$，如果对于 f 的任意子图 f'，均满足 $support(f,G)<support(f',G)$，则 f 为频繁闭图。

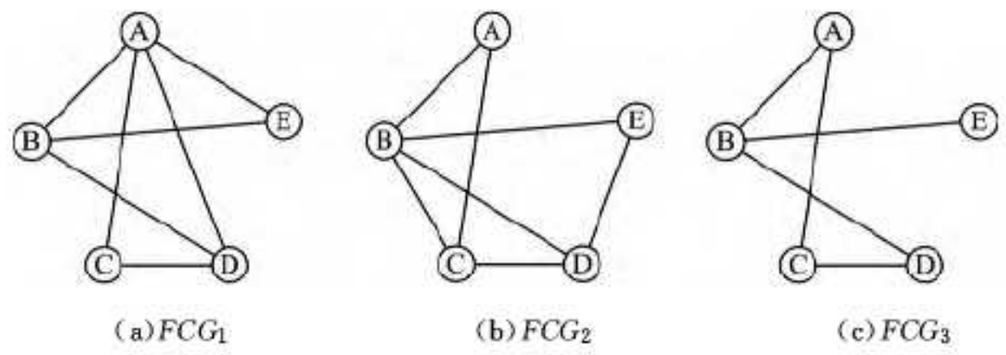
为了对频繁闭图有一个直观的认识，下面就以图 1 所示的动态网络为例进行说明。



(a)t_1 时刻　(b)t_2 时刻　(c)t_3 时刻

图 1　动态网络示例

将最小支持度设为 0.5，根据定义 4 和定义 5，可以分别得到示例动态网络的频繁子图集合和频繁闭图集合，其中频繁子图共 255 个，频繁闭图共 3 个，分别如图 2 所示。



(a)FCG_1　(b)FCG_2　(c)FCG_3

图 2　示例动态网络的频繁闭图

由图 2 可知，3 个频繁闭图的边数分别是 7、7、5，在频繁子图集合中属于较大者，两个最大的 7 边频繁子图都是频繁闭图，说明频繁闭图不仅能够在时间尺度上满足最小支持度，而且能够在空间尺度上覆盖动态网络的主要结构。

2.2　问题描述

动态网络的链路预测问题可以形式化描述为：给定一个动态网络 $G=\langle G_1,G_2,\cdots,G_T\rangle$，预测 t 时刻中的边集 E_t，其中 $t>T$。

本文所采取的链路预测方法的核心思想是，如果动态网络中的边 y 通常在边 x 出现后以较高的概率出现，则说明边 x 和边 y 在时间尺度上具有关联性，通过边 x 可以有效地预测边 y 的出现。

根据上述思想，链路预测的关键在于挖掘边与边之间的时序关联规则，最直接的方法是以边为单位挖掘关联规则，结合边 x 出现的时间以及边组 $\langle x,y\rangle$ 的时延，就可以推断边 y 出现的时间，这种链路预测方法主要存在以下两点不足：

1）如果考虑所有边的组合和可能的时间间隔，则计算全部 $I(x,y)$ 的时间复杂度为 $O(|V|^4T)$，无法处理大规模的动态网络；

2）很多边组在时间尺度上并没有关联性，因此计算整个边集空间的时间间隔会产生大量的噪声。

动态网络中每个网络快照都可以由若干个频繁闭图进行表示，频繁闭图可以近似表示一组具有统计显著性的边集，而且频繁闭图的数量比边数少很多，为了降低输入空间的规模，本文采取基于频繁闭图的链路预测方法。该方法将研究对象由原来的边集 E 替代为频繁闭图集合 FCG，其计算复杂度降为 $O(|FCG|^2T)$。相应地，由于 FCG 只能表征动态网络的骨干拓扑结构，因此动态网络链路预测问题就是预测 $t(t>T)$ 时刻的网络快照中骨干网络所包含的边集。

为便于度量任意子图序列 $\langle x,y\rangle$ 的时序关联性，本文引入子图时延的概念，如定义 6 所示。

定义 6（子图时延） 给定动态网络 $G=\langle G_1,G_2,\cdots,G_T\rangle$，假设网络快照 $G_m(1\leqslant m<T)$ 中包含子图 x 而不包含子图 y，如果子图 x 下一次出现的时刻为 $n(n>m+1)$，子图 y 出现的时刻为 $k(m<k<n)$，则子图 x 和子图 y 的时延 $I(x,y)=k-m$。

在子图 x 两个相邻出现时刻之间的时序中，子图 y 可能会多次出现，因此子图序列 $\langle x,y\rangle$ 的时延并不是一个固定值，而是服从一个时延分布。时延分布 $I(x,y)$ 具有以下两个性质：

1）如果 $x\neq y$，则 $I(x,y)\neq I(y,x)$；

2）$I(x,x)$ 有意义，表示子图 x 在动态网络中相邻两次出现的时间间隔。

为便于理解，下面就以图 2 所示动态网络中的子图序列 $\langle FCG_1,FCG_2\rangle$ 为例对时延分布进行说明。动态网络中每个时刻包含的频繁闭图集合分别为 $\{FCG_1,FCG_3\}$、$\{FCG_2,FCG_3\}$、$\{FCG_2,FCG_3\}$ 和 $\{FCG_1,FCG_2,FCG_3\}$，t_1 时刻中包含 FCG_1 且不包含 FCG_2，而 FCG_1 的下一个出现时刻为 t_4，在 t_1 与 t_4 时刻之间，FCG_2 的出现时刻为 t_2 和 t_3，根据子图时延的定义，$\langle FCG_1,FCG_2\rangle$ 的时延分布为 $\{1,2\}$。

在动态网络中，对于特定的子图序列 $\langle x,y\rangle$，每个时刻都可以计算一次子图时延分布，因此 $\langle x,y\rangle$ 的时延分布会随着时间不断变化。通过计算一段时间内的子图时延分布，可以得到特定子图时延出现的次数，如果 $\langle x,y\rangle$ 的某个时延出现的频率很高，则说明这一时延具有统计意义，可以在一定程度上反映子图 x、y 在时间尺度上的关联性，因而可以据此预测子图 y 出现的时刻。

所有子图序列的时延分布构成了时延分布矩阵 D，计算矩阵 D 是链路预测算法的核心。给定一段时间内的动态网络以及矩阵 D，就可以预测未来一段时间内的动态网络。

3　链路预测算法

3.1　算法流程

根据链路预测问题的形式化分析，可以将链路预测算法概括为如图 3 所示的框架。

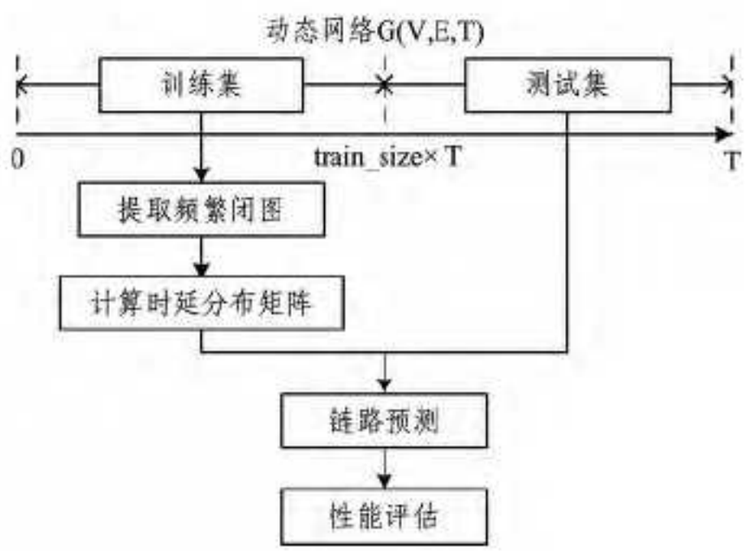


图 3　链路预测算法框架

由图 3 可知，链路预测算法主要分为 3 个阶段：提取频繁闭图、计算时延分布矩阵和链路预测。给定动态网络 $G=\langle G_1,G_2,\cdots,G_T\rangle$，算法按照时间轴将动态网络划分为训练集和测试集，前 $train\_size\times T$ 个网络快照作为训练集（$0<train\_size<1$），后 $(1-train\_size)\times T$ 个网络快照作为测试集。下面对算法中的 3 个阶段分别进行介绍。

(1)频繁闭图提取阶段

最直接的频繁子图挖掘方法就是穷举法，即对边集 E 穷尽各种组合，然后计算每一种边集组合的支持度，以判定每种组合是否为频繁子图。边集的组合总数最多可达 $2^{|E|}$，随着边数的增长，边集组合的数量呈指数级增长，因此这种方法的实用价值不大。Apriori 算法[8]是频繁项集挖掘中的经典算法，使用多轮迭代的方法逐步挖掘频繁项集，利用频繁项集的以下两条性质来降低搜索空间：

1)频繁项集的任何非空子集都是频繁的；

2)非频繁项集的任何超集都是非频繁的。

基于上述两条性质，Apriori 算法可以用于挖掘动态网络的频繁闭图集合。首先将训练集中的每个网络快照存储为 $|V|\times|V|$ 的下三角邻接矩阵，并将邻接矩阵序列化，形成长度为 $|V|^2$ 的向量，然后将 T 个时刻的向量组合成为 $|V|^2\times T$ 的二维矩阵，用于存储动态网络的所有边集 E。基于 Apriori 思想挖掘频繁闭图的实现方法如算法 1 所示。

算法 1　Apriori 频繁闭图挖掘算法

输入：动态网络 G=(V,E,T)，最小支持度 minsup

输出：频繁子图集合 FSG，频繁闭图集合 FCG

```
FSG_1={e|support(e,G)≥minsup,e∈E}
  for (k=1;|FSG_k|>1;k++)
    for (i=1:|FSG_k|)
      for (j=i+1:|FSG_k|)
        c=Candidates(FSG_k(i)∪FSG_k(j))
        if support(c,G)≥minsup
          FSG_{k+1}=FSG_{k+1}∪c
          if support(c)==support(FSG_k(i))‖support(c)==
            support(FSG_k(j))
            NCG_{k+1}=NCG_{k+1}∪c
      end
    end
  end
  FSG=∪_{i=1}^{k} FSG_i, NCG=∪_{i=2}^{k} NCG_i
  FCG=FSG-NCG
```

由算法 1 可知，Apriori 频繁闭图挖掘算法采取广度优先的候选子图遍历策略，首先从动态网络中提取所有的频繁边作为初始的候选子图，然后随着边数的递增，迭代式挖掘频繁子图，每一轮迭代后生成的频繁子图集合作为下一轮迭代的输入，直到生成的频繁子图边数小于 2 时，由于无法进一步组合形成具有更多边数的候选子图，最终输出所有的频繁子图和频繁闭图。

在每一轮的迭代中，Candidates 函数基于频繁子图的两条性质将边数为 k 的频繁边集进行组合，以生成边数为 $k+1$ 的候选边集；然后，对边数为 $k+1$ 的候选边集的支持度进行计数，如果支持度大于最小阈值，则可以加入边数为 $k+1$ 的频繁子图集合中；最后，对边数为 $k+1$ 的频繁子图集合进行剪枝，排除不符合频繁闭图定义的频繁子图。

(2)时延分布矩阵生成阶段

在提取到动态网络的频繁闭图集合 FCG 后，将训练集中的每个网络快照都表示为 FCG 的子集，然后在此基础上计算每一对频繁闭图的时延分布，得到时延分布矩阵 D，计算方法如算法 2 所示。

算法 2　时延分布矩阵的计算

输入：动态网络 G=(V,E,T)，训练集比例 train_size，频繁闭图集合 FCG，频率阈值 α

输出：时延分布矩阵 D

```
d=train_size×T
lagdist[|FCG|×d]=0
for i=1:|FCG|
  for k=1:d
    if FCG{i}⊂G{k}
      nexti=NextOccur(G(k+1:d),FCG{i})
      for j in (FCG-FCG{i})
        lag=AllOccur(G(k+1:nexti),FCG{j})
        if lag≠Φ
          lagdist(j,lag)++;
      end
    end
  end
  if lagdist(j,lag)>α
    D(i,:,:)=lagdist
end
```

在算法中，NextOccur 函数用于计算下一个时刻子图 i 的出现时刻 $nexti$，AllOccur 函数用于计算子图 j 在 $nexti$ 之前所有出现的时刻。通过遍历所有的频繁闭图和训练集的时长，可以得到任意子图序列之间的时延以及该时延出现的次数，如果特定时延的出现次数高于预设的频率阈值 α，则将该时延加入到矩阵 D 中相应子图序列所在时延分布中，最终输出大小为 $|FCG|\times|FCG|\times d$ 的时延分布矩阵 D 作为子图序列在时间尺度上的关联规则模型。

(3)链路预测阶段

链路预测的目标就是结合训练集中得到的时延分布矩阵以及测试集数据预测 $train\_size\times T$ 之后的动态网络，并求得预测结果的精确率和召回率，预测方法如算法 3 所示。

算法 3　链路预测算法

输入：动态网络 G=(V,E,T)，训练集比例 train_size，频繁闭图集合 FCG，时延分布矩阵 D，可信度阈值 β

输出：动态网络 PG(V,E′,T-d)，精确率 precision，召回率 recall

```
d=train_size×T
R[|FCG|×|FCG|×d]=1
for i=1:|FCG|
```

```
occur(i)=AllOccur(G(d:T),FCG{i})
if occur(i)≠Φ
  for j in (FCG−FCG{i})
    if R(i,j,lag)>β
      occur(j)=occur(i)+ D(i,j,lag)
      PG( FCG{j},occurj)=1
        if FCG{j} not in G{occurj}
          update(R(i,j,lag))
  end
end
inter(d:T)=G(d:T) & PG(d:T)
precision=sum(inter(d:T))/sum(PG(d:T))
recall=sum(inter(d:T))/sum(G(d:T))
```

为了避免不可靠的预测，算法为矩阵 D 中的每个时延分配可信度 R，初始化为 1。$D(i,j,lag)$ 表示子图序列 $\langle FCG_i, FCG_j\rangle$ 的时延 lag 出现的次数，如果经过时延 lag 后 FCG_j 并没有出现在网络快照中，则认为该时延 lag 是错误的，将其错误次数记为 $W[i,j,lag]$，则 lag 的可信度定义为：

$$R(i,j,lag)=\frac{D(i,j,lag)}{D(i,j,lag)+W(i,j,lag)} \tag{1}$$

将测试集中频繁闭图 FCG_i 的所有出现时间记为 $occur(i)$，任取另一个频繁闭图 FCG_j，如果时延可信度 $R(i,j,lag)$ 高于预设的可信度阈值 β，则推断 FCG_j 出现在 $occur(i)+lag$ 时刻的网络快照中，然后对比测试集中相应时刻的网络快照，如果 FCG_j 没有出现在该网络快照中，则更新 $R(i,j,lag)$。通过遍历频繁闭图集合和时延分布矩阵，最终输出预测得到的动态网络 $PG(V,E',T-d)$。

3.2 评价指标

衡量链路预测算法性能的指标有很多，如 AUC、Ranking Score 等，这些指标通常被应用于以边为单位的预测算法中。为了提高计算效率，上述链路预测算法以频繁闭图为单位进行预测，预测得到的每个网络快照都是由若干个频繁闭图的边集构成的，无法全面覆盖动态网络中的所有边集 E，因此，在输出的预测结果 $PG(V,E',T-d)$ 中，边集 $E'\subset E$。

为了有效地分析本文提出的链路预测算法的性能，使用精确率 $precision$ 和召回率 $recall$ 的加权调和平均数 $F\text{-}score$ 作为评价指标。首先，根据公式分别计算预测结果的 $precision$ 和 $recall$。

$$precision=\frac{\sum_{i=1}^{|E|}\sum_{t=d}^{T}G(V,E,T-d)\&PG(V,E',T-d)}{\sum_{i=1}^{|E'|}\sum_{t=d}^{T}PG(V,E',T-d)} \tag{2}$$

$$recall=\frac{\sum_{i=1}^{|E|}\sum_{t=d}^{T}G(V,E,T-d)\&PG(V,E',T-d)}{\sum_{i=1}^{|E|}\sum_{t=d}^{T}G(V,E,T-d)} \tag{3}$$

其中，精确率 $precision$ 用于衡量预测得到的动态网络 PG 中正确边集占 E' 的比例，召回率 $recall$ 用于衡量 PG 中正确边集占 E 的比例。$F\text{-}score$ 的计算方法如式(4)所示。

$$F\text{-}score=\frac{(\sigma^2+1)\,precision\times recall}{\sigma^2\,precision+recall} \tag{4}$$

通过调节参数 σ，可以改变 $precision$ 和 $recall$ 在 $F\text{-}score$ 中的权重，$\sigma=1$ 时 $F\text{-}score$ 是标准型。为了使 $F\text{-}score$ 指标适用于基于频繁闭图的链路预测算法，应该适当降低 $recall$ 的权重，在下面的实验中，统一取 $\sigma=0.5$。

4 实验和结果

为了分析链路预测算法的性能，下面就以特定国家的 AS 级 Internet 在一段时间内的动态网络为例进行分析。

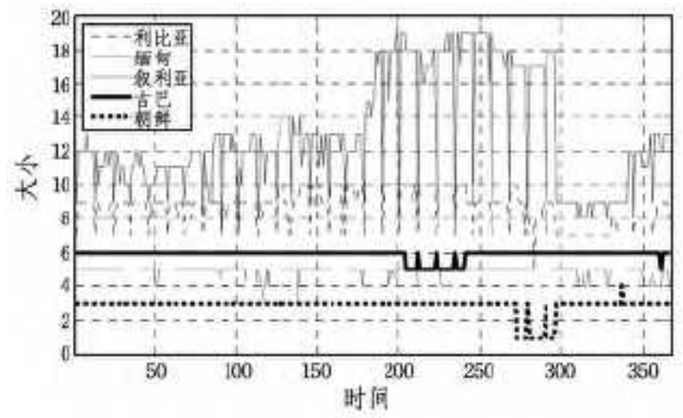
4.1 数据集

本文以 RouteViews 项目[9]提供的 RIB 表数据集作为研究对象，为了进行对比研究，分别选取两个时间跨度、时间粒度各不相同的数据集。表 1 所列为每个数据集的摘要信息。

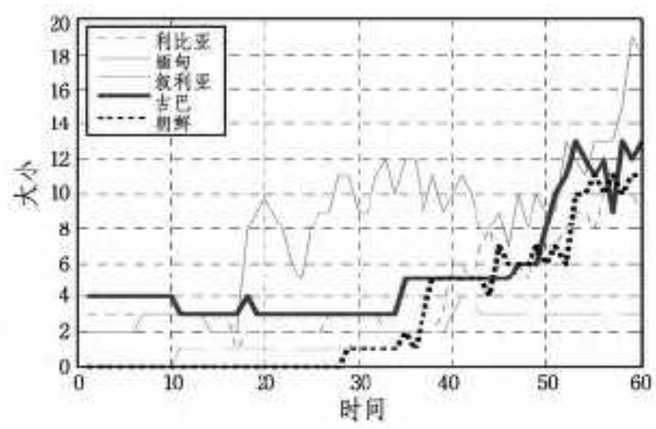
表 1 RIB 表数据集信息

	数据集 1	数据集 2
时间粒度	日	月
起始时间	2014 年 1 月 1 日	2010 年 1 月 1 日
结束时间	2014 年 12 月 31 日	2015 年 1 月 1 日
大小	11.37GB	2.23GB
国家	利比亚、缅甸、叙利亚、古巴、索马里	利比亚、缅甸、朝鲜、突尼斯、索马里

为了在分析算法性能的同时验证算法的有效性，选取部分网络基础设施比较脆弱的国家，将这些国家的所有 AS 以及与其直接相连的 AS 构成的 AS 级 Internet 作为研究对象。数据集 1 的时间跨度短、粒度小，分别选取表中各国 AS 级 Internet 进行分析，得到各国网络大小的演化趋势，如图 4(a) 所示；数据集 2 的时间跨度长、粒度大，分析得到表中各个国家网络大小的演化趋势，如图 4(b) 所示。



(a) 数据集 1



(b) 数据集 2

图 4 各国 AS 级 Internet 动态网络大小演化趋势

由图 4(a) 可知，在 2014 年内，各国网络大小的演化曲线始终在平均值附近起伏波动，适于分析算法在波动型动态网络中的性能；由图 4(b) 可知，在 2010 年至 2014 年的 5 年时间里，各国网络大小呈现明显的增长趋势，适于分析链路预测算法在增长型动态网络中的性能。

4.2 算法性能分析

首先分析算法在波动型动态网络中的性能。由链路预测算法的输入可知，影响算法性能的参数主要包括支持度阈值 $minsup$ 和频率阈值 α，下面就分析两个参数对算法 $precision$、$recall$ 和 $F\text{-}score$ 3 个指标的影响。

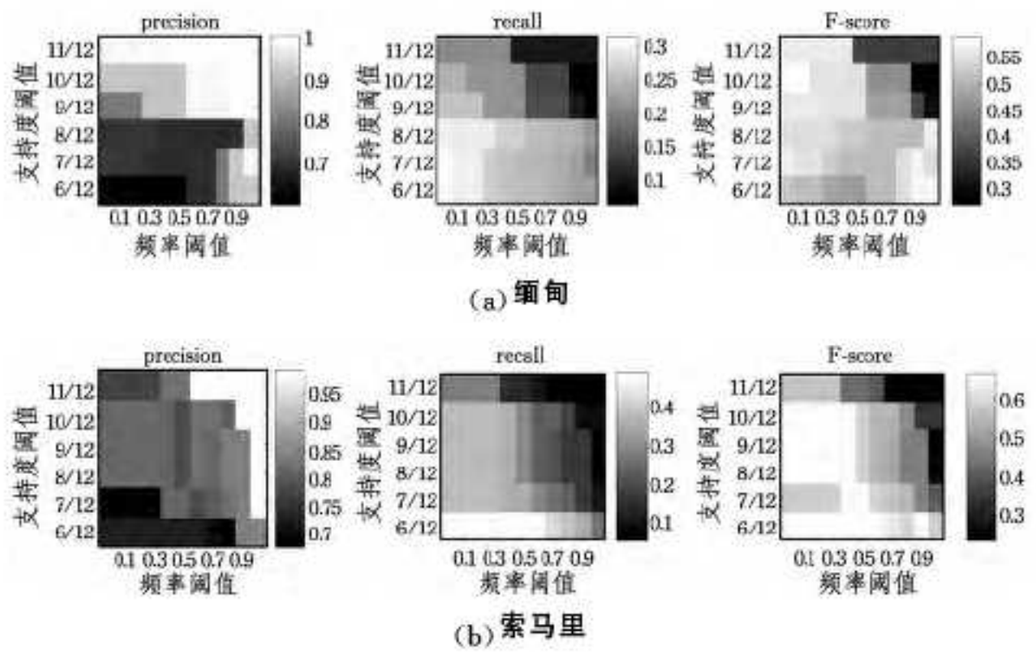
将算法分别运用于数据集 1 中各国 AS 级 Internet 动态网络中，取前 6 个月的动态网络作为训练集，后 6 个月的动态

网络作为测试集，训练集主要用于生成频繁闭图集合和时延分布矩阵。各个国家预测结果的平均 *precision* 和 *recall* 值如表 2 所列。

表 2　波动型动态网络中算法各项评价指标的平均值(%)

国家	precision	recall
利比亚	81.7	28.7
缅甸	83.85	21.7
叙利亚	85.3	25.31
古巴	80.91	29.65
索马里	82.24	28.35

从表 2 中的平均值看出，预测结果的 *precision* 通常比较高，每个国家的 *precision* 平均值均在 80% 以上；而 *recall* 取值范围相对较低，说明链路预测算法得到的动态网络虽然无法覆盖实际动态网络中的大部分边集，但是能够达到很高的预测精度，因而可以有效地预测动态网络中的骨干网络结构。以缅甸和索马里动态网络为例，在取不同支持度阈值和频率阈值的情况下，算法的 *precision*、*recall* 和 *F-score* 的取值分布如图 5 所示。



(a)缅甸

(b)索马里

图 5　算法在不同国家动态网络中的各项指标

由图 5 可知，尽管算法在不同国家动态网络中的 *F-score* 取值分布不尽相同，但是从 *precision* 和 *recall* 的取值情况看，二者符合相似的分布规律。

算法的 *precision* 与支持度阈值和频率阈值均呈正相关，这是因为支持度阈值越高，算法得到的频繁闭图集合的表征能力越强，在测试集中出现频繁闭图的概率就越高；而频率阈值越高，过滤后的频繁闭图之间的时序相关性更强，在此基础上预测得到的动态网络的精确率肯定更高。与 *precision* 相反，*recall* 与支持度阈值和频率阈值均呈负相关，支持度阈值越高，从训练集中提取得到的频繁闭图数量越少，而频率阈值越高，算法会剔除更多时序相关性较弱的频繁闭图，因此二者的提高都会使预测得到的动态网络边集占测试集中动态网络边集的比例降低。正是由于 *precision* 和 *recall* 取值分布的矛盾性，使得 *F-score* 的取值与支持度阈值和频率阈值均呈非线性相关，无法在特定的支持度阈值和频繁阈值下取得最大值。比如，在缅甸的动态网络中，*F-score* 在支持度阈值和频率阈值为 $(8/12, 0.1)$ 时取最大值；而在索马里的动态网络中，*F-score* 在支持度阈值和频率阈值为 $(6/12, 0.9)$ 时取最大值。

接下来分析算法在增长型动态网络中的性能。将算法运用于数据集 2 的各国动态网络中，取前 30 个月的动态网络作为训练集，后 30 个月作为测试集，得到各评价指标的取值范围，如表 3 所列。

表 3　增长型动态网络中算法各项评价指标的取值范围

国家	precision	recall	F-score
利比亚	NaN	[0,0.0226]	NaN
缅甸	NaN	[0,0]	NaN
朝鲜	NaN	[0,0]	NaN
突尼斯	NaN	[0,0]	NaN
索马里	NaN	[0,0.123]	NaN

由表 3 可知，算法在增长型动态网络中的性能很差，不具有可操作性，这主要是因为在算法的频繁闭图提取阶段，训练集是动态网络中边集最小的 30 个网络快照，从中很难提取可用于链路预测的频繁闭图集合，而且数据集的时间粒度太大，这在客观上降低了频繁闭图之间的时序相关性。

结束语　为了实现利用已知动态网络预测未知动态网络，本文提出了一个基于频繁闭图的链路预测算法，其核心假设是两个频繁闭图在时间尺度上具有关联性，通过挖掘已知动态网络中频繁闭图之间的时序相关性预测下一段时间的动态网络结构。将该算法分别运用于不同时间尺度的大规模 Internet 动态网络中，实验结果表明，算法能够以很高的精确率预测波动型动态网络中未知时刻的骨干网络，但是不适用于增长型动态网络结构的预测。

本文提出的链路预测算法在异常事件推断、网络动态性分析和网络动力学等方面具有广阔的应用前景。在下一步的研究中，一方面可以尝试将训练集窗口化，不断提取新的频繁闭图；另一方面，可以在识别频繁闭图时引入松弛参数，从而增加频繁闭图的数量，以利于提高链路预测算法的召回率；此外，研究适用于增长型动态网络的链路预测算法也是今后研究的重点。

参 考 文 献

[1] 傅颖斌，陈羽中. 基于链路预测的微博用户关系分析[J]. 计算机科学，2014，41(2)：201-205

[2] Zhu J，Hong J，Hughes J G. Using markov chains for link prediction in adaptive web sites[M]// Soft-Ware 2002：Computing in an Imperfect World. Springer Berlin Heidelberg，2002：60-73

[3] O'Madadhain J，Hutchins J，Smyth P. Prediction and ranking algorithms for event-based network data[J]. ACM SIGKDD Explorations Newsletter，2005，7(2)：23-30

[4] Liben-Nowell D，Kleinberg J. The link-prediction problem for social networks[J]. Journal of the American society for information science and technology，2007，58(7)：1019-1031

[5] Lahiri M，Berger-Wolf T Y. Structure prediction in temporal networks using frequent subgraphs[C]// IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007). IEEE，2007：35-42

[6] Clauset A，Moore C，Newman M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature，2008，453(7191)：98-101

[7] Yan X，Han J. CloseGraph：mining closed frequent graph patterns[C]//Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2003：286-295

[8] Inokuchi A，Washio T，Motoda H. An apriori-based algorithm for mining frequent substructures from graph data[M]//Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg，2000：13-23

[9] University of Oregon Route Views project [EB/OL]. http://www.routeviews.org/