*Four Degrees of Separation*

*——信息检索阅读报告*

*10152510239*

*邱梦华*

**一、背景和实验基础**

Frigyes Karinthy,在他 1929创作的Láncszemek”(“Chains”)中提出，任意两个人都能通过至少六个人联系，即六度分离理论。Stanley Milgram也在他的著名实验中，要求人们只通过直接熟人把明信片发给一个固定的收件人，明信片的转手次数平均数介于4.4和5.7之间。

本文的实验的基础实验为米尔格拉姆的著名的“小世界”实验。但是，米尔格拉姆测量的是社交网络上的路由路径的平均长度，并且这等同于平均距离的上界。许多人基于这个实验在各种大型网络上进行了尝试。

本文Facebook和米兰大学共同宣布了他们关于六度分离理论的新研究成果：他们已经确定世界上任何两个独立的人之间平均所间隔的人数为4.74，使用的数据为脸谱网的活跃用户（大概七亿两千一百万的用户），这是第一个世界级规模的社会网络图距离的计算。

本文使用的主要工具是印度海得拉巴的第二十届万维网会议上，Sebastiano提出的一个研究非常大的图的距离分布的新工具： HyperANF。在以前的图压缩工作和扩散计算思想的基础上，新工具使得可以精确地研究图的距离分布大于以前可能的数量级。

**二、Definitions and Tools**

**（1）我们的讨论假设为：**

假设时间t的边m（t）的数目与节点n（t）的数目有关：



其中A是常数，通常位于区间（1，2）

**（2） 一些定义**

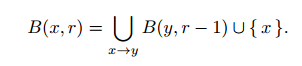
①邻域函数 N G (t)

图G的邻域函数 N G (t)表示从x到y最多需要t步的节点对（x，y），t属于N。它提供了关于每个节点周围的“average ball”有多快的数据。从邻域函数可以得出距离分布（在距离T上的可达对之间）。

② HyperANF算法和HyperLogLog计数器

在本文中，我们使用基于扩散HyperANF算法，它能够快速地逼近非常大的图的邻域函数。使用WebGraph来表示一个被压缩但快速可访问的图表。

HyperANF中，球的节点X附近的半径R满足

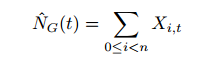


且由于我们可以使用图的顺序扫描，递增地计算每个B（x，r）。但是在扫描过程中，我们需要随机地访问集合B（x，r-1） 。由于所需的空间太大，不能保存在主存储器中。HyperANF使用HyperLogLog计数器以近似的方式表示这些集合，可以被理解为可以可靠地回答关于大小问题的字典。每个计数器是由许多小的寄存器组成的。寄存器跟踪一个良好的哈希函数的值的最大值M，它应用于节点序列的元素：序列中的不同元素的数量与2M成比例。将流分成多个子流，每个子流由不同的寄存器分析。然后通过适当地聚合来自每个寄存器的估计来计算结果。

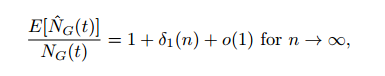
面临的主要性能挑战是如何快速计算与球的结合相关联的 HyperLogLog计数器，每一个都由一个HyperLogLog计数器来表示：HyperANF使用基于字级并行性的算法，该算法使计算速度非常快，并且经过精心设计。实现利用多核架构，在内核的数量线性加速。HyperANF的另一个重要特征是它使用收缩的方法来避免重新计算不需要的球。迭代过程中更改的方法是能够计算整个距离分布的基础，避免了先前方法所使用的没有可证明的精度的任意终止条件。

**（3）理论误差界**

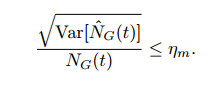
在第n次迭代中运行一个HynANF的结果是T中的邻域函数的估计，我们可以把它看作一个随机变量。

其中每个xi，t是计数在T步中由节点i达到的节点的HyperLogLog计数器（n是图的节点数）。当使计数器使用M个寄存器时,每个xi，t具有保证的相对标准偏差ηM≤1.06/ √ m.

第t次迭代HyperANF中NG（t)的的渐近几乎无偏估计，即



其中δ1与〔定理1〕相同（ |δ 1 (x)| <5 \*10 −5当 m ≥ 16），此外，N G (t)有一个相对标准偏差不大于X i ’s,



特别是,除了两个最大的脸谱网图，其中我们使用m＝32（εm＝0:187）,其他的我们使用m＝64（εm＝0:1325）.重复运行，以获得一致的相对标准偏差的所有图表,但是，邻域函数的相对误差变为距离分布的绝对误差。实际上，通过最后值除以邻域函数值来获得累积分布函数的简单行为是非线性的，并且在估计中引入了偏差。为了减少偏差，并提供我们的测量的标准误差的估计，我们使用 jackknif，一种经典的非参数方法来评估数据样本的任意统计，这在实践中是非常有效的。

**三、实验流程**

本文分析的图表是2011五月活跃的脸谱网用户（最近28天内登录的用户）的图，将重点放在合理地代表实际个体的账户上。我们将实验扩展到区域和时间两个方向。因此，我们分析整个脸谱网图（FB），美国子图（US），意大利子图（IT）和瑞典（SE）子图，我们还分析了意大利和瑞典的图形（ITSE）的组合，以检查是否结合两个区域遥远的网络可以显著改变平均距离。设置

1. **设置，运行**

在24核机上进行计算，内存为72GIB，磁盘空间为1TIB。

由于大部分图的大小，必须通过内存映射来访问，所以我们决定将所有的计数器b（x，r-1）和b（x，r）存储在主存储器中，以避免异位 I/O。在当前的整个脸谱网图上，HyperANF的运行使用了32个寄存器，所以SPA计数器的CE是大约27 GIB。

且根据数据分析，2006年9月脸谱网向非大学生开放：我们的数据显示订阅量急剧上升。特别是，2007年1月1日的IT和SE子图高度断开，甚至脸谱网本身也断开连接，但是我们计算的所有数据在2009之后稳定（具有小的振荡），基本上所有的对都可以到达。因此，我们认为2007和2008的数据用于观察脸谱网的进化，但我们不认为它们代表了底层人类社会链接结构。

1. **分布，平均度和密度**

图2显示了当前图的概率质量函数，我们可以区分区域图集中在距离四，整个脸谱网图集中在距离五附近。此外，IT和SE的分布具有明显小于ITSE和US的概率集中在距离五上的概率质量。方差数据（表7和图4）表明，分布变得非常集中。

表4显示了我们所考虑的所有图的平均程度的相对快速增长。越多的用户加入网络，就发现了更多的现有的友情关系。在图6中，我们展示了同一数据的对数空间标度比例图：用我们所处理的小的点集很难得出可靠的结论，但是我们并不总是观察幂律行为。相反，网络的密度降低。在图5中，我们绘出了图的密度与节点的数量有一些初始的交替行为，但在更完整的网络（FB和US）稀疏化的趋势是非常明显的。地理集中反而增加密度，我们可以看到对应于我们的区域图明确地按地理集中排序的线，FB图在最低位置。

1. **平均距离**

图3和表6中显示的平均距离结果可知，脸谱网图上的平均距离是4.74。92%的可达对个体在距离五或更少。研究分析发现平均距离实际上取决于用户的地理接近性，而不仅仅取决于网络的实际大小。而在脸谱网增长最快的年份，我们的图表显示了平均距离的快速下降，但是现在则比较稳定。因此，我们在工作中看到了小世界现象：连接用户的弧的一小部分，但是平均距离较低。并且在表11中，形象化了“分离度”概念，即平均距离上限中的可达对的百分比。

**（4）Spid，直径**

SPID是距离分布离散的指标σ 2 /µ ，一些人提出SPID可以作为社交网络的“webbiness”的度量，SPID大于1的网络应该被认为是“网络状”，而SPID小于1的网络应该被认为是“适当社会”，因此网络被认为是否是适当的社会取决于其距离分布是否分散。

对于社交网络来说测量距离分布的分散的spid 小于1（欠分散），但对于网络图则大于1（过分散）。SPID背后的直觉是“适当的社交”网络强烈支持短连接，而在Web中长连接并不少见。但是研究结果平均距离与SPID之间的相关性一般不大。然而，如果我们考虑与单个网络相关联的点，则平均距离与SPID之间似乎有一定的相关性，方差在相同的方向上变化：较小的平均距离（这意味着较大的SPID）与较小的方差相关。图8显示了我们分析的图中平均距离和方差之间的轻度相关性：随着网络越来越紧密，它的距离分布也变得更加集中。

对于脸谱网的研究有结论：这个世界比我们想象的要小，当前脸谱网图的平均距离是4:74。此外，图的SPID是0:09，证实了适当的社交网络的SPID远远低于1。

HyperANF不能提供关于直径的精确结果，但是运行的步数必然是图的直径的下界。在表10中，我们将这些最大观测值作为区域图和整体脸谱网图之间不同的下界，有人比世界上的单个国家更明显地“相距甚远”。为了证实它，我们也决定了直接计算精确直径的问题，实现了一个高度并行版本的iFUB（迭代条纹上界）算法并扩展为无向图。

**（5）准确性**

从距离分布导出的数据很难获得强的理论界。问题是，当从邻域函数传递到距离分布时，由于计算平均距离或其他统计数据时距离分布达到非常小的值存在产生严重误差风险，相对误差界变为绝对误差界。另一方面，派生数据的分布极其集中。但是，由经验解释结果也是出乎意料的正确。由于计数器，我们仍然认为任何基于扩散的算法有效的。

**四、总结**

我们从多个角度研究了有史以来最大的电子社交网络Facebook（包括7亿2100万个活跃的用户及其690亿个友情链接）。

首先我们证实了，分层标记传播是一个强大的通过增加其节点的社会网络的局部性的范例。我们已经能够以每比特56%的信息理论下限来压缩US图以及其他更小的社交网络。

然后，我们进行了我们进行了当前能达到的最大规模的米尔格拉姆式实验，使用了 HyperANF最终得到了脸谱网的平均距离为4.74，也就是3.74“分离度”,脸谱网的SPID是0:09远远低于一个，符合社交网络的预期。

总的来说，这些结果有助于描绘脸谱网社交图。正如预期的那样，它是一个小世界图，在许多节点对之间具有短路径，得出如题的四度分离理论。这个研究将六度理论改写为四度理论，即当你想到世界上另一个人时，你朋友的朋友认识他朋友的朋友。