专题：统计物理和复杂系统

网络直播平台数据挖掘与行为分析综述*

郭淑慧 吕欣†

(国防科技大学系统工程学院, 长沙 410073)

(2019 年 11 月 22 日收到; 2020 年 3 月 20 日收到修改稿)

随着移动通信和互联网技术的不断发展, 网络直播逐渐成为了新媒体环境下人们青睐的在线娱乐和信息传播方式. 目前广泛应用于课堂教学、真人秀、电竞赛事、品牌营销等方面. 数百万主播与数亿计观众的活跃加入和互动, 产生了丰富的在线人群行为活动数据, 为开展大规模人群行为动力学、平台内容推荐与检测、在线社群演化等研究提供了丰富的实验场景. 本文通过梳理国内外网络直播平台数据挖掘与行为分析的相关研究文献, 分析了直播平台负载水平、观众行为、主播行为以及社群网络的特征和变化规律, 并对直播平台中大规模人群行为表现出的时空规律和重尾效应进行了总结. 直播平台中各种社群网络的形成和演化机制、内容推荐与检测等是未来网络直播领域研究的发展趋势.

关键词： 直播平台, 用户行为, 社群网络, 数据挖掘

PACS： 89.70.–a, 89.75.–k, 89.75.Kd **DOI：** 10.7498/aps.69.20191776

1 引 言

网络直播是一种新型信息交流方式, 可以让观众收看到主播所处场景中正在进行的音、视频实况, 同时观众可以通过打赏或发表评论的方式与主播进行互动, 相对于传统信息传播媒体来说有着互动性强、时空适应性强等优势. 网络直播平台则是由公司或组织管理的供主播发布直播以及观众收看直播的网站. 随着互联网经济的发展, 网络直播日益火爆, 斗鱼 TV、虎牙 TV、抖音等网络直播平台在市场中异军突起, 以超低的门槛吸引了大量主播和观众, 截至 2019 年 6 月, 我国直播用户规模已达 4.33 亿[1].

网络直播目前的应用领域比较广泛, 除了应用于娱乐性的真人秀、电竞赛事之外, 还有课堂教学[2,3]、品牌营销[4,5]、传统文化与工艺技术传承[6,7]、政务会议与庭审过程公开[8,9]等方面. 不同领域的网络直播和观众都会产生大量交互数据, 这些数据一方面可以用于挖掘直播平台的负载变化模式和用户参与及交互的内在机制, 探究用户行为和偏好, 进而对相应情境下的大规模人群行为进行模式分析和规律挖掘[10–17]. 另一方面还可以基于直播平台负载的测量结果及从中挖掘的用户行为的特征和偏好, 提升网络直播平台内容推荐和内容检测水平[18–22].

本文从国内外网络直播平台用户行为数据挖掘的研究入手, 对直播平台负载水平、观众行为、主播行为以及社群网络的特征和变化规律进行梳理和总结, 并讨论网络直播平台研究在当前面临的问题和未来的研究方向.

2 直播平台负载研究

网络直播是通过网络直播平台进行实时信息传输的新媒体形式. 随着网络的发展和普及, 网络

* 国家自然科学基金 (批准号: 82041020, 71771213, 91846301, 71790615) 和湖南省科技计划项目 (批准号: 2017RS3040, 2018JJ1034) 资助的课题.

† 通信作者. E-mail: xin.lu@flowminder.org

直播作为一种学习、娱乐的便捷资源被人们越来越广泛地使用. 直播平台负载的水平差异反映了直播平台用户的分布规律和使用偏好, 对直播平台的负载研究可以从整体上把握直播平台的资源消耗与服务使用情况, 对直播平台优化资源配置、提供经济稳定的负载支持有指导意义.

目前关于平台负载研究主要是通过统计直播平台的运行负载, 挖掘负载水平产生规律性差异的时间因素、空间因素以及其他影响因素, 总结直播平台情境下的大规模人群行为偏好和行为特征.

2.1 时序特征

受时间节律的影响, 人类行为会在诸多方面不同程度地体现出日内效应 (diurnal effect)、周内效应 (weekly effect) 等时序规律, 如金融市场的流动性[23]、人类的情绪积极程度[24]、反应灵敏度[25]、器官工作机能[26] 等都会在一天内不同时段表现出显著差异, 股市收益率和波动还存在明显的周内效应[27]. 目前对直播平台的负载研究大部分集中于从系统带宽、主播规模、观众数量、打赏额和评论量等方面的时序变化规律中挖掘直播平台负载的日内效应、周内效应和长期规律等时序特征.

Veloso 等[11] 最早根据巴西某网络电视直播平台的网站日志对负载的时序特征进行了研究. 在 2002 年为期 28 天的统计中, 用户的访问模式显示出了明显的日内效应和周内效应, 昼夜模式造成凌晨 4:00—11:00 在线观众数量偏低, 峰值和谷值分别在 3:00 和 9:00 附近取得; 双休日的平均观众数量明显高于工作日. 尽管该直播平台的用户规模超过 69 万、覆盖 65 个国家, 但受当时的网络发展水平限制, 平台使用的带宽峰值仅仅为 80 Mbps.

随着 Twitch, YouTube Live 等专门化网络直播平台的兴起, Kaytoue 等[28] 根据 2011 年末 Twitch 平台的直播间数量和在线观众数量变化对直播平台负载的周内效应进行分析, 发现 Twitch 平台的双休日负载明显高于工作日 (后续学者[29,30] 也得出了一致结论), 原因是 Twitch 平台的主要直播内容是电子竞技, 而大型电子竞技的竞赛通常在双休日举行. Pires 和 Simon[31] 对比了 Twitch 平台和 YouTube Live 平台在 2014 年的系统带宽和直播间数量发现, 两个平台的带宽峰值都超过了 1 Tbps, 但 Twitch 平台的带宽水平更高, 峰值超过了 1.6 Tbps. 两个平台在直播间的数量在双休日都明显高于工作日, 但 Twitch 平台和 YouTube Live 的日内负载峰值分别在 5:00 和 18:00 附近取得, 而且 Twitch 平台在日内和周内负载变化模式的敏感度均低于 YouTube Live. 原因是 Twitch 平台的开放时间较早, 用户在全球范围内的覆盖范围更广, 减弱了昼夜更替造成的时序差异.

近年来逐渐出现了对国内直播平台负载的研究, Zhu 等[32] 通过收集国内直播平台斗鱼 TV 在 2016 年 12 月为期 14 天的直播数据, 发现观众数量和主播数量在一天中显示出几乎一致的变化规律, 都在 21:00—8:00 减少, 8:00—21:00 增加, 在晚上 9:00—10:00 达到最高水平. Wang 等[33] 通过分析 2016 年 9 月起为期 124 天内的斗鱼 TV 主播开播数量、观众打赏总额和评论量来挖掘斗鱼 TV 平台负载的日内效应和周日效应, 从观众的打赏总额、评论量和主播直播次数、直播时长分别展现观众和主播在一天之中的活跃程度变化趋势. 结果发现观众和主播的活动都表现出很强的昼夜规律, 并且高度同步, 任意两个数据系列之间的皮尔逊相关系数都高于 0.85. 但观众活动的高峰时段出现在 23:00—24:00, 主播最活跃的时段是 20:00—22:00, 说明观众的活跃时间存在一定的时滞现象, 与之前的研究结论[32] 略有差异.

总体上说, 国内外直播平台负载在一天之中都呈现“倒 N 型”[34], 直播平台负载具有明显的“日内效应”[25], 负载水平在一天中呈现降低-升高-降低的循环模式 (如图 1 所示).

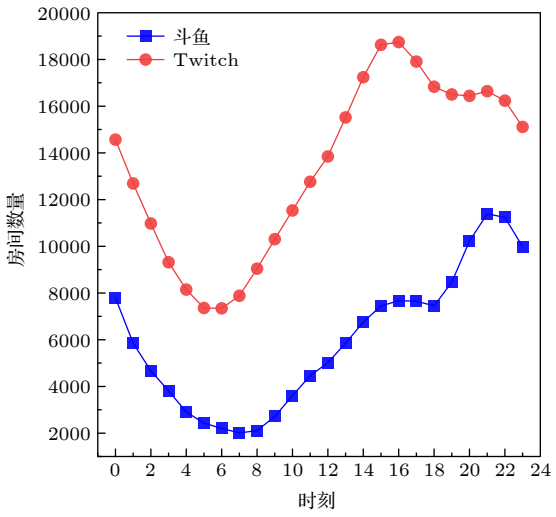

图 1 国内外直播平台负载日内效应[30,32]

Fig. 1. The diurnal effects of domestic and foreign live streaming workloads[30,32].

表 1 不同直播平台的负载时序特征

Table 1. The workload changes of different live streaming platforms.

直播平台	年份	日内效应	周内效应
Anonymous[11]	2002	3, 19时观众数量最多	双休日观众数量更多
Twitch[30]	2015	17, 0时直播间最多	双休日直播间数量显著增多
YouNow[35]	2015	直播间和观众最多: 22时	双休日直播间和观众数量更多
斗鱼[32]	2016	直播间和观众最多: 21—22时	—
斗鱼[33]	2018	直播间最多: 20—21时, 观众最多: 23—24时	双休日直播间和观众数量更多

但由于直播平台的直播类型各有侧重, 而且直播平台随着时间在逐渐发展, 即使是同一直播平台在不同的统计期内表现的平台负载时序规律也不完全相同. 多个直播平台、不同时期的负载变化时序特征见表 1.

2.2 空间分布特征

除了分析直播平台负载的时序特征之外, 部分研究通过分析主播、观众和直播平台服务器的位置对直播平台负载产生的影响来挖掘直播平台负载的地理分布特征, 进而对直播负载的资源分配及优化进行指导.

Veloso 等[11] 对早期网络电视直播平台的观众数量和观看次数在所覆盖区域间的数量分布进行了统计分析, 发现观众数量和观看次数在划分的地理区块之间的分布形式都近似 Zipf 分布[36]:

$$Z(r) = Ar^{-\alpha}, \tag{1}$$

其中参数r代表地理区块按照观众数量或观看次数降序排列的排名, $Z(r)$则代表排名为r的区域中的观众数量或者观看次数, α的取值分别为 1.29 和 1.49, 展现了早期网络电视直播的观众在地理分布上的不均匀现象. 与之相对的, Li 等[37] 统计分析了 PPTV 直播频道的观众地理分布熵 (viewer geographical entropy) 的累积分布函数, 发现大多数频道的观众地理分布熵超过 0.8, 显示了现代网络直播平台观众观看者在地理位置上均匀分布. 观众地理分布熵形式为

$$e_k = \frac{1}{\log N}\left(-\sum_{i=1}^{N} p_{ki}\log p_{ki}\right), \tag{2}$$

e_k是直播间k的观众地理分布熵, 其中N是直播间全部观众所覆盖区域的数量, p_{ki}代表直播间k在区域i的观众数量占全部观众数量的比例.

Kaytoue 等[28] 通过统计分析 Twitch 平台的主播在不同时区的分布, 表明平台中的大多数主播都来自北美、欧洲和东亚, 与 Twitch 平台的服务器集中布置在北美、欧洲和亚洲的分布规律[38] 相符合, 反映了直播平台负载的地理特征对直播平台服务器设置的指导意义.

Yan 等[39] 对比了用户生成视频、短视频和直播视频三种平台的城市、郊区及整个区域的移动网络用户的观众地理分布熵, 结果显示无论市区、郊区还是整个区域, 用户生成视频的熵都高于其他服务, 表明观看直播的用户比观看用户生成视频的用户在空间分布上更不均匀, 主要原因是直播内容通常耗费的流量更多, 用户更倾向于在固定的场所使用 WiFi 或宽带网络而不是移动流量来观看直播.

2.3 其他影响因素

除了时空对直播平台负载的影响之外, 少量学者对主要直播电子竞技内容的直播平台 Twitch、斗鱼 TV 的负载水平是否受到大型电子竞技竞赛项目直播的影响进行了研究. Kaytoue 等[28] 通过观察 Twitch 平台在 2011 年 12 月 29 日到 2012 年 1 月 9 日每天的观众数量变化情况, 发现在直播一些重要的电子竞技比赛时, 观众数量会出现明显增加, 说明了电子竞技竞赛项目直播对 Twitch 平台的负载有强烈的刺激作用. 类似地, Deng 等[12] 统计了 Twitch 平台电子竞技竞赛项目直播吸引的观众占整个平台观众的比例, 结果显示某些热门的电子竞技竞赛项目直播所产生的观众数量能占直播平台全部观众的 30% 以上, 即使是不太流行的电子竞技赛事的直播也能吸引大量观众, 峰值超过全平台观众 10%. 但 Wang 等[33] 在对斗鱼 TV 观众评论数以及打赏额在 2016 年为期 124 天的统计期中的变化中却并未发现重大赛事对平台负载产生的显著影响, 原因可能是斗鱼 TV 存在大部分娱乐类直播, 受电子竞技竞赛项目直播的影响并不明显.

3 观众行为分析

直播平台内可能出现以下的一种或几种观众行为: 选择直播间进行观看、切换或退出直播间、在直播间中评论或打赏、以及观众观看直播而引发的行为 (如被主播引导购买商品). 众多学者对直播平台观众行为中的观看规律进行分析和建模, 研究观众各种行为以及背后的心理, 对于理解用户参与网络直播的原因、提升用户体验、为用户提供更有价值的网络直播服务有重要的决策价值.

已有研究中对观众的观看规律主要从观众的观看次数与时长、频道选择与切换、观众评论与打赏等方面入手, 从观众的观看记录中提取直播平台中观众的各种观看行为, 挖掘其中观众的偏好及心理动机, 进一步开展直播平台观众的行为动力学建模和社群网络演化研究.

3.1 观看次数与时长

众多研究结果表明, 直播平台内观众的观看时长、观看次数呈现一定的重尾效应, 即直播平台中存在大量观看次数很少、观看时间非常短的用户, 但同时还有极少量用户观看直播的次数很高、观看时长相对非常长.

Veloso 等 [11] 于 2002 年的研究结果显示早期直播平台中观众观看次数分布近似 Zipf 分布, 其中参数 r 代表的是按照观看次数降序排列的观众排名, $Z(r)$ 则代表排名为 r 的观众的观看次数, 参数 α 的取值为 0.719; Li 等 [37] 于 2016 年发现 PPTV 平台内观众观看次数分布更符合互补 Weibull 分布而不是幂律分布, 累计概率分布函数形式为

$$P(X \geqslant x) = \exp[-(x/x_0)^c], \tag{3}$$

其中参数 c 为拉伸因子, x_0 为常数参数.

Sripanidkulchai 等 [40] 则于 2004 年对早期直播平台观看时长分布进行了探索, 结果显示不同直播间内观众的观看时长分布均近似 Zipf 分布, 其中参数 r 代表的是按照观看时长降序排列的观众排名, $Z(r)$ 则代表排名为 r 的观众的观看时长, 参数 α 的取值在 0.7—2.0 之间. 而 Tang 等 [41] 于 2006 年发现 CCTV 多个直播频道的观众观看时长的概率密度函数形式符合对数正态分布:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left[\frac{-(\ln x - \mu)^2}{2\sigma^2}\right], \tag{4}$$

其中 $f(x)$ 是观看时长为 x 分钟的概率密度函数值, μ 和 σ 的取值均在 4—5 和 1—2 之间. Li 等 [37] 于 2016 年对 PPTV 中观众观看时长分布的概率密度函数进行了研究, 结果显示观众观看时长的概率密度函数形式为混合指数分布:

$$f(x) = \sum_{i=1}^{n} \alpha_i \frac{1}{\mu_i}\exp\left(-\frac{1}{\mu_i}x\right), \tag{5}$$

其中 $f(x)$ 是观看时长为 x 分钟的概率密度函数值, μ_i 和 α_i 是第 i 个指数分布的均值和权重, $\sum_{i=1}^{n}\alpha_i = 1$.

特别地, Tang 等 [41] 通过分析观众观看直播的时长记录, 发现观众已经观看直播的时长与继续保持观看的时长存在显著的正相关关系, 即观众如果已经花费了比其他观众长的时间观看直播, 则会更倾向于比其他观众花费更长的时间继续观看直播.

3.2 频道选择与切换行为规律

如果在观看直播的过程中发生网络故障、主播关闭直播、不感兴趣等情况, 观众就有可能对当前直播间进行重新连线、切换到其他直播间或者直接退出直播平台. 目前已有相关研究对观众在观看直播中的重连、切换、退出等进行统计建模与系统分析, 展现直播平台内观众流动的动态过程和内在机理.

Li 等 [37] 对客观原因造成直播中断时的观众重连行为进行了研究. 由于观看出现中断的原因可能是网络连接失败等客观问题, 也可能单纯是由于观众的兴趣发生变化而主动退出直播间, 所以作者首先定义了由客观问题造成中断的直播段特征, 是观众在某直播间内发出观看请求之后的一小段时间之内对该直播间重复发出了观看请求. 进一步分别统计了观众对直播中断次数的容忍程度和放弃观看的概率分布, 结果显示随着直播中断次数的增加, 观众放弃的概率递增, 但增幅在逐次减小. 在移动网络下观看直播的观众在直播中断时的耐心程度比在 WiFi 或者宽带环境下的更高, 在遇到 2 次连续的直播中断时, 放弃观看的概率小于 50%, 甚至在某些情况下观众能忍受 10 次连续的直播中断.

Nascimento 等 [42] 对 Twitch 平台中的观众切换行为进行了建模分析. 作者首先定义了直播间共存在三种状态, 分别是直播中、直播即将结束和直播结束 (如图 2 所示). 由于主播下播之后直播间内的观众并不会被强制清空, 所以会出现主播已经下

播但观众数量不为 0 的情况. 直播间三种状态中的“直播即将结束”包含了主播即将关闭直播以及关闭直播后观众数量仍保持一定水平时期, “直播结束”指的是直播间内观众数量非常低甚至为 0 的状态.

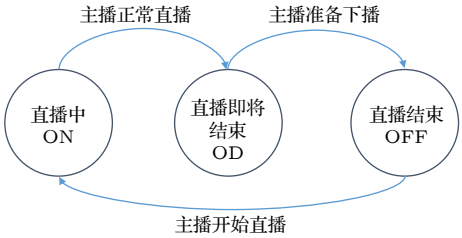

图 2 直播间状态演化图[42]

Fig. 2. The graph of live streaming channel's state dynamics[42].

由直播间的三种状态任意组合可以归纳出观众切换行为共包含 9 种类型, 通过统计观众不同类型切换行为的比例发现观众在频道之间的切换行为中, 大约 20% 的 ON-ON 切换 (即从一个正在直播的直播间切换到另一个正在直播的直播间) 和 30% 的 OFF-OFF 切换 (即从一个未开播的直播间切换到另一个未开播的直播间) 持续不到 1 min, 表明观众进入直播间时并不事先知道他们想观看的内容. 77% 的切换行为是 ON-X(X 可能是 ON 或者 OFF), 这说明大部分的直播间切换是由于观众对切换前直播间的内容不满意.通过对比观众在 ON-ON 切换前后的观看时长的分布, 发现观众在切换后的直播间会观看更长的时间, 同样验证了大部分观众是主动退出了之前不满意的直播间. 由于观众对直播间选择的随意性较强、满意度较低, 部分研究 [18–20] 设计了针对直播平台的内容推荐算法及系统, 为观众选择直播频道进行个性化推荐.

Li 等[37] 对直播节目中观众的进入和退出过程进行了建模和分析. 通过统计直播过程中观众加入和离开的速率变化, 发现在直播节目开始之前的一小段时间内会出现观众大量进入和立刻退出的现象, 而且观众的加入和离开速率在很大程度上依赖于某些子事件的发生. 考虑到以上因素, 作者使用高斯径向基函数 (RBFs) 之和来拟合直播过程中观众加入和离开行为, 较低的 RMSE 值表明模型与实际过程相符. 形如

$$\lambda(t)=\sum_{i=1}^{m}a_i\exp\left[\frac{-(t-T_i)^2}{2{\sigma_i}^2}\right], \tag{6}$$

其中 T_i 是第 i 个子事件发生的时刻, a_i 是子事件 i 的影响权重.

3.3 评论与打赏行为分析

观众向直播间赠送虚拟礼物被称之为“打赏”, 已有研究对国内直播平台内的打赏金额分布规律进行挖掘. Zhao 等[43] 统计了映客平台内某些主播收到的打赏, 结果显示观众打赏的金额服从长尾分布. 也就是说, 极少比例的观众贡献了大部分打赏, 贡献排名前 20 名的观众的打赏额占所有观众打赏的 90% 以上. 相似地, Zhu 等[32] 通过统计斗鱼 TV 中打赏额在主播之间的分布情况, 发现打赏额在主播之间分布的幂律性, 几个最受欢迎的主播拥有非常高数量的打赏, 其他频道分享的很少, 而且仅 2.7% 的打赏观众贡献了 80.2% 的打赏额. 兰荣亨等[44] 则根据观众的观看、评论和打赏记录对观众群体行为进行特征构建, 成功对不同特征的观众群体进行了聚类区分.

Wang 等[33] 对打赏行为的时间规律进行分析, 发现不论是第一次打赏出现的时间还是打赏间隔时间的分布形式都高度符合 Weibull 分布, 其累计分布函数形为

$$F(t)=\Pr[X\leqslant t]=1-\mathrm{e}^{-(t/\lambda)^k}, \tag{7}$$

其中 λ 和 k 是分布的尺度和形状因子. 拟合结果显示两个分布的形状因子 k 均小于 1, 即说明直播平台中主播已经等待打赏的时间越长, 那么后续打赏到来所需的等待时间越长.

观众评论是观众利用文字和表情符号在直播间中进行交流的一种方式, 目前对观众评论的研究主要是对评论的情感、特征、观众交互进行分析. Poyane[45] 对 Twitch 平台部分直播 Dota2 的直播间内的观众评论文本数据进行了情感分析, 发现随着直播间观众规模的增加, 观众评论的消极色彩会相应增强. 类似地, Nematzadeh 等[46] 也发现随着观众数量的增加, 评论区会由正常对话向过载的、不和谐对话转变.

Olejniczak[47] 对 Twitch 平台观众的评论内容从语句特征上进行了分析. 发现观众更倾向于使用大量的表情符号和重复信息来表达态度, 使用新颖的词汇和独特的表情符号来力求与众不同. 由于评论区只显示最新的几条评论内容, 评论长度会随着观众数量增加而缩短. Li 等[33] 发现了观众评论与

打赏的周内模式存在很强的相关性 (皮尔森相关系数超过 0.85), 且评论与视频内容同步性很强, 提出了根据评论情感标注直播亮点的算法[48]. 周钰淇[22]则提出了根据评论内容对直播内容是否合法进行检测的深度学习算法.

4 主播行为分析

直播平台中的主播行为研究主要集中在主播的直播次数、直播时长以及流行度排名与预测三方面. 通过分析主播群体独特的行为模式, 挖掘直播平台内主播直播规律和活动特征, 对进一步开展直播平台内大规模人群行为分析和研究、优化直播平台系统建设有重要意义.

4.1 直播次数

国内外直播平台的主播直播次数普遍呈现出一定程度的幂律分布特点, 即直播次数较少的主播占了很大部分, 直播次数多的主播占比很少[14,32,35,42].

对国外直播平台的主播直播次数规律的研究中, Stohr 等[35]挖掘 Younow 平台的主播直播数据发现, 超过 40% 的主播只直播了一次, 约 10% 的主播在一周内直播了 7 次以上. 这表明, 有一小部分高度活跃主播愿意每天直播多次, 而大多数主播仅进行少量的直播. 类似地, Nascimento 等[42]发现 Twitch 平台的少部分专业主播团队直播的次数达到每天 19 次, 而大量主播 (40%—50%) 每天直播次数不超过 1 次, 与 Jia 等[14]的结论基本一致.

对国内直播平台的主播直播次数规律的研究中, Zhu 等[32]通过统计 14 天内斗鱼 TV 主播开播天数的分布, 发现约 63% 的主播每周直播的天数不超过 1 天, 只有 14% 的主播在统计期中至少直播了一半的时间. 大多数主播并不经常直播, 但整个平台每天约有 4% 的主播会进行直播.

4.2 直播时长

众多研究表明主播在直播时长方面表现出重尾分布规律. Zhu 等[32]通过收集斗鱼 TV 主播在为期 14 天统计期内的所有开播记录来统计主播直播时长分布, 发现 70% 以上的直播时长都小于 200 min, 但存在极少比例 (小于 1%) 的主播直播时长达到 1000 min. 统计结果显示斗鱼 TV 主播的直播时长中位数是 90 min, 比 Twitch 平台的 45 min[28]更长, 原因是相对于 Twitch 平台主要针对游戏内容进行直播, 斗鱼 TV 的直播类型更多样, 许多直播间播放已经制作好的视频或大型活动, 提高了直播的持续时间. 类似地, 对 Younow[35]和 Twitch 平台[14,30,42,49]的主播直播时长特征的研究结果均显示主播直播时长分布呈现重尾效应.

研究中通常以直播平台内所有主播直播时长的中位数作为衡量该直播平台主播直播时长的指标, 由于直播内容和针对观众等方面的差异, 各个直播平台的直播时长中位数不尽相同. 即使是相同的直播平台, 不同直播类别的直播时长中位数也会有一定的变化. 部分结论总结如表 2 所列.

表 2 各个网络直播平台的直播时长中位数

Table 2. Median live streaming duration of each live streaming platform.

文献	直播平台	采集年份	中位数
[35]	YouNow	2015	16 min
[42]	Twitch(StarCraftII的较大型直播间)	2013—2014	3.7 h
[30]	Twitch(点播的视频)	2015	8 min
[49]	Twitch	2014	150 min
[28]	Twitch	2011—2012	95 min
[32]	斗鱼	2016	90 min

4.3 流行度排名和预测

主播的“流行度”指的是主播吸引观众的能力. 通过某些衡量指标对主播吸引力进行排名, 排名越靠前则说明主播吸引观众的能力越强、主播流行度越高. 这一排名在体现主播在直播平台中的地位和水平的同时也反映出了观众的访问模式. 目前的研究中通常以粉丝或观众数量、打赏金额、评论数量作为衡量指标来对直播平台主播的流行度进行排序.

大量对主播流行度排名的研究显示, 主播流行度存在一定的重尾分布规律[12,14,30,35,40,43,45,49,50]. Pires 等[31,50]研究发现 Twitch 平台的观众数量分布符合 Zipf 分布, 且参数 α 的值在 1.0—1.5 之间变动, 说明大量观众在很少几个直播间中聚集, 少数主播吸引了绝大部分观众. Stohr 等[35]对 Younow 平台观众数量分布的研究也得出了类似的结论. 但 Zhang 和 Liu[49]则发现 Twitch 平台的观众数量分布形式不是标准的幂律分布. 由于著名主播通过直播吸引了极大比例的观众观看, 观众数量分布的尾部出现了明显的下降, 更符合 Gamma分布或者 Weibull 分布的特点, 与 Wang 等[33]对斗鱼 TV

内评论数量和打赏金额分布形式一致. Arnett 等[51]则对主播在社交平台上的公开活动对主播流行度是否产生影响进行了研究, 并没有发现主播的观众和粉丝数量变动与在社交平台上的活动存在统计学上显著的相关性.

对主播流行度预测的研究中, Kaytoue 等[28]分析线上内容发布后短期和长期的流行度相关性, 提出线性回归模型, 以此来通过前期观众数较准确地预测后期观众数. 还提出了一种流行度的定义 (不仅仅只比较在线人数, 还考虑上线时间的早晚等因素), 并以此对主播进行新的排序. 基于主播流行度预测线性回归模型, Netzorg 等[52]提出了基于主播行为的主播未来流行度预测模型, 发现主播的努力行为 (如发布更多直播、定期直播、在其他社交媒体账号上发布直播信息等) 在提升主播流行度方面是有效的, 而且职业主播比业余主播更受欢迎. 类似地, Szabo 等[53]用浏览次数代表视频的流行度, 从前期数据预测视频未来长期的流行度. Zhu 等[32]对直播间出现的总观众数和礼物总价值进行线性相关分析, 计算得出直播间内观众总数和礼物总价相关系数是 0.6421 ($p < 0.001$). Jia 等[14]计算得出直播间在线人数和主播直播次数的相关系数也较高, 即直播次数多的主播流行度可能更大.

5 社群网络分析

网络直播吸引了大量的主播与观众参与, 直播平台中大规模人群交互形成了很多独具特色的社群现象. 通过统计分析直播用户的使用特征, 识别和发现直播平台内的社群及社群网络, 进一步分析直播社群网络的节点属性、结构特征以及形成、演化过程, 开展对直播平台大规模人群参与、流动及交互的规律挖掘和动力学研究, 对信息传播、网络营销、舆情监测引导等策略的制定等都有参考和指导意义.

5.1 社群发现与成员识别

直播平台具有的社交属性使得其中存在大规模人群的交互关系, 从而形成了多种类型的用户关系复杂网络. 而直播平台用户网络的节点属性、结构和形成演变机制则体现了直播平台情景下大规模人群活动的交互特征和选择偏好. 部分研究对用户观看、关注、评论、打赏等关系网络中的社区发现[54]方法进行了探索, 通过识别用户关系网络中相似用户形成的社区为直播平台用户关系网络发展动态的研究奠定基础.

Churchill 和 Xu[55]于 2016 年发表了首个对直播平台用户社区的研究并提出了社区发现和成员识别算法. 该社区发现算法是通过可视化主播共享观众关系网络实现主播社区识别. 作者首先收集了游戏直播平台 Twitch 的主播信息, 包括主播直播的游戏类型及粉丝列表. 进一步地, 作者构建了以主播为节点, 共享观众关系为边的主播关系网络, 其中节点大小代表了主播拥有的粉丝数量, 颜色代表主播的直播游戏类型, 主播之间共享的粉丝数量越多, 那么连边越粗、节点之间的距离越短. 通过可视化主播关系网络, 根据节点颜色和距离的分布直观分析主播之间联系的紧密程度, 实现主播关系网络中的社区规模和结构的识别.

作者提出的社区成员识别算法则是根据主播与主播之间的关注关系, 自动识别出社区成员. 作者首先人工挑选出实际属于 Twitch 平台三大主流社区[56]的四位主播作为种子节点, 种子节点的关注者中粉丝量在 28000 以上的则被程序自动判定与该主播所属同一社区, 从而实现了对主播所在社区成员的识别和发现. 识别结果与 Gephi 中的模块化识别结果基本一致, 说明了社区发现算法的有效性. 类似地, Lykousas 等[57]通过设定违规用户作为种子节点, 从 Live.me 平台和 Loops Live 平台的观看关系网络中自动判定用户是否违规, 实现了违规用户所在社区的成员发现.

5.2 社群网络节点重要性

前文总结了直播平台中观众数量、评论量和打赏金额在主播间的分布普遍呈现出重尾效应, 说明以主播为节点, 以观众观看、评论或打赏为节点重要性衡量指标的网络中, 存在少量中心节点, 它们在整个主播社群网络中的地位和重要性非常高. 但由于网络直播发展的时间较短, 目前尚未出现对直播平台主播社群网络的节点性质进行专门分析的文献.

在观众与主播共同形成的社群网络中, 由于观众对主播有天然的选择权利, 社群中的观众成员对主播的喜爱和认可促使主播成为了社群的意见领袖和核心, 研究普遍认为主播促进了整个社群建立

和发展, 对社群的发展方向起决定作用. William 等[58] 对 Twitch 平台的主播和观众进行访谈, 发现直播间的社群氛围折射了主播的品质和态度, 即主播成为了整个直播间的意见领袖. 而且直播间内的核心成员发挥吸引其他参与者、促进互动以及缓和聊天的重要作用以建立社群, 也即社群中的意见领袖促进了整个社群建立和发展. 庄庆玲和周丽[59] 以斗鱼游戏主播为例研究了弹幕式互动直播平台主播和观众之间形成的社群, 发现在直播间的互动中主播会成为意见领袖, 与追随者也就是观众基本属于同一阶层, 拥有共同的兴趣但意见领袖对该领域有更全面和深入的了解, 依靠平台进行有偿信息交流.

5.3 社群网络结构特征

研究复杂系统内的社群网络结构有助于理解或预测系统的表现[60], 少量研究对直播社群网络的结构特征进行了描述, 通过社群网络的结构性质反应直播社群的交互特点, 以解释直播平台中的社群表现.

Churchill 和 Xu[55] 对 Twitch 平台主播社群的网络结构性质进行了研究. 发现主播直播的游戏相同或相似, 那么他们拥有的粉丝相似度也很高, 体现了观众对直播类型的偏好性. 而且通过识别 Twitch 平台中三大主流社区的成员, 发现社区规模大小与直播难度相关, 成为该类主播的难度越大则该类主播数量越少. Lykousas 等[57] 通过对直播平台违规用户所形成的观看网络结构进行统计分析, 对网络的平均度、密度和相互性进行分析发现, 违规用户网络的互惠性很差 ($\text{reciprocity} < 0.15$), 即违规用户社群网络中节点与节点之间在互惠互利方面的表现并不好.

6 讨论与展望

综上所述, 网络视频直播用户行为挖掘的国内外相关研究目前已取得一定的进展. 从大量相关文献发现, 直播平台负载、观众行为、主播行为和社群网络是本领域的研究重点. 其中, 直播平台负载水平的变化模式体现出了明显的日内效应和周内效应, 大规模人类行为在直播平台中体现出明显的重尾特征, 如观众观看次数与时长、打赏额和评论量、主播直播次数和时长、吸引观众的能力等分布均从不同程度上符合重尾分布. 说明直播平台内人群分布的异质性很强, 可以据此对直播平台的经营模式如用户 (观众、主播) 激励模式、虚拟礼物打赏机制等进一步优化. 网络直播平台中大规模人群交互形成了多种用户社群网络, 识别用户社群和分析社群网络的结构和属性对优化直播平台的发展和应用有重要意义.

由于网络直播以及相关研究发展的时间尚且较短, 对网络直播的研究广度和深度有待进一步探索. 从研究主题发展轨迹和当前研究重点可以预测, 挖掘直播平台中各种社群网络的形成和演化机制、设计针对直播平台的内容推荐和检测算法等是网络直播领域研究的未来发展趋势.

参考文献

[1] CNNIC http://www.cac.gov.cn/2019-08/30/c_1124938750.htm [2019-8-30]

[2] Huang S, Hu H 2002 *Proceedings International Symposium on Multimedia Software Engineering* Taiwan, December 11–13, 2000 p411

[3] Abdous M, Yoshimura M 2010 *Comput. Educ.* **55** 733

[4] Cao M X 2019 *Mod. Mark.* **(10)** 82 (in Chinese) [曹孟熙 2019 现代营销 (经营版) **(10)** 82]

[5] Lu Z, Xia H, Heo S, Wigdor D 2018 *Proceedings of the 2018 CHI conference on human factors in computing systems* Montréal, Canada, April 21–26, 2018 p1

[6] Yan M 2019 *Southeast. Commun.* **(5)** 66 (in Chinese) [阎敏 2019 东南传播 **(5)** 66]

[7] Dong D L 2019 *Course Educ. Res.* **(41)** 24 (in Chinese) [董弟林 2019 课程教育研究 **(41)** 24]

[8] Bao X 2009 *Gov. Leg. Inst.* **(22)** 51 (in Chinese) [保旭 2009 政府法制 **(22)** 51]

[9] Guo B, Zhang H Y 2008 *Inherit* **(22)** 18 (in Chinese) [郭波, 张会永 2008 传承 **(22)** 18]

[10] Borges A, Gomes P, Nacif J, Mantini R, Almeida J M, Campos S 2012 *Comput. Commmun.* **35** 1004

[11] Veloso E, Almeida V, Meira W, Bestavros A, Jin S 2002 *Proceedings of the 2 nd ACM SIGCOMM Workshop on Internet measurment* Marseille France, November 6–8, 2002 p117

[12] Deng J, Cuadrado F, Tyson G, Uhlig S 2015 *2015 International Workshop on Network and Systems Support for Games (NetGames)* Zagreb, Croatia, December 3–4, 2015 p1

[13] Haimson O L, Tang J C 2017 *Proceedings of the 2017 CHI conference on human factors in computing systems* Colorado, USA, May 6–7, 2017 p48

[14] Jia A L, Shen S, Epema D H, Iosup A 2016 *Acm T. Multim. Comput.* **12** 47

[15] Bonald T, Massoulié L, Mathieu F, Perino D, Twigg A 2008 *ACM SIGMETRICS Performance Evaluation Review* MD, USA, June 2–6, 2008 p325

[16] Chang H, Jamin S, Wang W 2009 *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement* Illinois USA, November 4–6, 2009 p417

[17] Fallica B, Lu Y, Kuipers F, Kooij R, Mieghem P V 2008 *The Second International Conference on Next Generation Mobile Applications, Services, and Technologies* Cardiff, Hnited Kingdom, September 16–19, 2008 p501

[18] Yang T W, Shih W Y, Huang J L, Ting W C, Liu P C 2013 *Conference on Technologies and Applications of Artificial Intelligence* Taiwan, December 6–8, 2013 p188

[19] Liu Y W, Lin C Y, Huang J L 2015 *2015 IEEE International Conference on Consumer Electronics-Taiwan* Taiwan, June 6–8, 2015 p118

[20] Lin C Y, Chen H S 2019 *Multimedia Tools Appl.* **78** 1999

[21] Zhang L 2019 *Mark. China* **(Z2)** 56 (in Chinese) [张璐 2019 成功营销 **(Z2)** 56]

[22] Zhou Y Q 2018 *M.S. Thesis* (Nanjing: Nanjing University of Posts and Telecommunications) (in Chinese) (in Chinese) [周钰淇 2018 硕士学位论文 (南京: 南京邮电大学)]

[23] Han D, Wang C, Yue H Y 2006 *Beijing Hangkong Hangtian Daxue Xuebao* 5 (in Chinese) [韩冬, 王春峰, 岳慧煜 2006 北京航空航天大学学报 (社会科学版) 5]

[24] Pink D H 2019 *When: The Scientific Secrets of Perfect Timing* (New York: Penguin Press) pp15–20

[25] Hines C B 2004 *J. Cathol. Educ.* **7** 390

[26] Bernard T, Giacomoni M, Gavarry O, Seymat M, Falgairette G 1997 *Eur. J. Appl. Physiol. Occup. Physiol.* **77** 133

[27] Chen X B, Zhang Z C 2008 *Chin. J. Manage. Sci.* **(4)** 44 (in Chinese) [陈雄兵, 张宗成 2008 中国管理科学 **(4)** 44]

[28] Kaytoue M, Silva A, Cerf L, Meira Jr W, Raïssi C 2012 *Proceedings of the 21st International Conference on World Wide Web* Lyon France, April 16-20, 2012 p1181

[29] Farrington D J, Muesch N M https://web.wpi.edu/Pubs/E-project/Available/E-project-031915-220004/unrestricted/Analysis_of_the_Characteristics_and_Content_of_Twitch.tv_Live-streaming.pdf [2015-3]

[30] Claypool M, Farrington D, Muesch N 2015 *2015 IEEE Games Entertainment Media Conference (GEM)* Toronto, Canada, October 14–16, 2015 p1

[31] Pires K, Simon G 2015 *Proceedings of the 6th ACM multimedia systems conference* Oregon, USA, March 18–20, 2015 p225

[32] Zhu Z H, Yang Z, Dai Y F 2017 *International Conference on Social Computing and Social Media* Vancouver, Canada, July 9–14, 2017 p274

[33] Wang X, Tian Y, Lan R, Yang W, Zhang X 2018 *IEEE T. Circ. Syst. Vid.* **29** 3454

[34] Peng S L, Wu N N, Zhao G Q 2014 *AMM* **675–677** 1810

[35] Stohr D, Li T, Wilk S, Santini S, Effelsberg W 2015 *2015 IEEE 40 th Local Computer Networks Conference Workshops (LCN Workshops)* Florida, USA, October 26–29, 2015 p673

[36] Newman M E J 2005 *Contemp. Phys.* **46** 323

[37] Li Z Y, Kaafar M A, Salamatian K, Xie G G 2016 *IEEE Trans. Circuits. Syst. Video* **27** 2675

[38] Deng J, Tyson G, Cuadrado F, Uhlig S 2017 *International Conference on Passive and Active Network Measurement* Sydney, Australia, March 30–31, 2017 p60

[39] Yan H, Lin T H, Zeng M, Wu J, Huang J X, Li Y, Jin D P 2017 *GLOBECOM 2017 : IEEE Global Communications Conference (GLOBECOM)* Singapore City, Singapore, December 4–8, 2017 p1

[40] Sripanidkulchai K, Maggs B, Zhang H 2004 *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement* Sicily, Italy, October 25–27, 2004 p41

[41] Tang Y, Sun L, Luo J, Zhong Y 2006 *7th Pacific-Rim Conference on Multimedia* Hangzhou, China, November 2–4, 2006 p175

[42] Nascimento G, Ribeiro M, Cerf L, Cesário N, Kaytoue M, Raïssi C, Vasconcelos T, Meira W 2014 *9th Latin American Web Congress* Minas Gerais, Brazil, October 22–24, 2014 p1

[43] Zhao J, Ma M, Gong W, Zhang L, Zhu Y, Liu J 2017 *2017 IEEE/ACM 25 th International Symposium on Quality of Service (IWQoS)* Vilanova i la Geltrú, Spain, June 14–16, 2017 p1

[44] Lan R H, Zhu G, Yang W, Tian Y, Zhu M 2019 *Comput. Syst. Appl.* **28** 69 (in Chinese) [兰荣亨, 朱格, 杨文, 田野, 朱明 2019 计算机系统应用 **28** 69]

[45] Poyane R 2018 *Proceedings of the 22nd International Academic Mindtrek Conference* Tampere, Finland, October 10–11, 2018 p262

[46] Nematzadeh A, Ciampaglia G L, Ahn Y Y, Flammini A 2016 *R. Soc. Open Sci.* **6** 191412

[47] Olejniczak J 2018 *Redefining Community in Intercultural Context* Bucharest, Romania, May 11–15, 2018 p329

[48] Lan R H, Hu Y H, Zhu G, Tian Y, Zhu M 2019 *Comput. Syst. Appl.* **28** 219 (in Chinese) [兰荣亨, 胡雨晗, 朱格, 田野, 朱明 2019 计算机系统应用 **28** 219]

[49] Zhang C, Liu J 2015 *Proceedings of the 25 th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video* Oregon, USA, March 18–20, 2015 p55

[50] Pires K, Simon G 2014 *Proceedings of the 2014 Workshop on Design, Quality and Deployment of Adaptive Video Streaming* Sydney, Australia, December 2, 2014 p13

[51] Arnett L, Netzorg R, Chaintreau A, Wu E 2019 *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* Glasgow, Scotland Uk, May 4–9, 2019 pLBW1211

[52] Netzorg R, Arnett L, Chaintreau A, Wu E 2018 arXiv: 1812.03379 [cs.SI]

[53] Szabo G, Huberman B A 2008 *Commun. ACM* **53** 80

[54] Clauset A, Newman M E J, Moore C 2004 *Phys. Rev. E* **70** 6

[55] Churchill B C, Xu W 2016 *2016 IEEE International Conferences on Big data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom)* Georgia, USA, October 8–10, 2016 p223

[56] Smith T, Obrist M, Wright P 2013 *Proceedings of the 11th European Conference on Interactive TV and Video* Como Italy, June, 2013 p131

[57] Lykousas N, Gómez V, Patsakis C 2018 *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* Barcelona, Spain, August 28–31, 2018 p375

[58] Hamilton W A, Garretson O, Kerne A 2014 *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems* Toronto, Canada, Aprial 26–May 1, 2014 p1315

[59] Zhuang Q L, Zhou L 2018 *J. News Res.* **10** 32 (in Chinese) [庄庆玲, 周丽 2018 新闻研究导刊 **10** 32]

[60] Newman M E J 2003 *SIAM Rev.* **45** 167

SPECIAL TOPIC—Statistical physics and complex systems

Live streaming: Data mining and behavior analysis*

Guo Shu-Hui Lu Xin †

(*College of Systems Engineering, National University of Defense Technology, Changsha* 410073, *China*)

(Received 22 November 2019; revised manuscript received 20 March 2020)

Abstract

With the rapid development of mobile communication and Internet technologies, online live streaming has gradually become popular for information communication and entertainment in the new media environment. Live streaming has been widely used in teaching, reality show, E-sports games and events, brand marketing and other aspects. With the active participation of millions of streamers and hundreds of millions of viewers, massive online crowd behavior activity data are generated, which offers rich experimental scenarios for large-scale crowd behavior dynamics research, live streaming channel recommendation and online community evolution. In this paper, we summarize the relevant research literature of live streaming, and review current studies from a comprehensive list of aspects: workload pattern, viewers and streamers behavior, community network discovery and analysis, etc. We summarize the temporal and spatial patterns of live streaming platform workload, heavy tailed effect of large-scale crowd behavior in live streaming platform, etc. We believe that the future work on live streaming can be directed in the examination of formation and evolution mechanism of various community networks formed by large-scale users, as well as the recommendation and detection of live streaming content.

Keywords: live streaming platform, human behavior, community network, data mining

PACS: 89.70.–a, 89.75.–k, 89.75.Kd **DOI:** 10.7498/aps.69.20191776

* Project supported by the National Natural Science Foundation of China (Grant Nos. 82041020, 71771213, 91846301, 71790615) and the Science and Technology Program of Hunan Province, China (Grant Nos. 2017RS3040, 2018JJ1034).

† Corresponding author. E-mail: xin.lu@flowminder.org