利用流行病学模型预测2022—23年科技行业裁员

RICHARD VALE

Abstract. 2022年和2023年，科技行业和初创企业行业的许多大大小小的公司都裁掉了异常多的员工。我们希望在不依赖经济预测的情况下，预测这段裁员时期何时可能结束。我们观察到，截至2023年3月31日的一组裁员数据，呈现出SIR（易感者—感染者—移除者）模型的噪声观测模式。我们使用Kr¨oger和Schlickeiser [KS20]提出的SIR模型解析解，对数据进行拟合。根据拟合模型，我们估计每周裁员人数将在2023年底左右恢复到正常水平。

arXiv:2305.05210v1 [stat.AP] 9 May 2023

1. Introduction

1.1. 2022-23 [SW23] [HML+23]期间，许多科技行业员工遭到裁员。多个在线信息聚合平台收集了科技行业裁员的自报数据，其中包括 `layoffstracker.com`、`trueup.io` 和 `layoffs.fyi`。本文旨在估计这段异常高裁员率持续期何时可能结束。

1.2. 由于公司裁员的决定往往取决于经济环境，预测未来裁员最自然的方法就是采用经济预测。然而，经济预测素来并不可靠 [OEC14]，因此，寻找一种不依赖经济预测的方法也颇有意义。

1.3. 在本研究中，我们采用这样一种观点，即裁员主要是一种社会现象。在2022年12月的一次访谈中 [dW22], ，杰弗里·普费弗教授指出，当前持续发生的科技行业裁员是一种“社会传染”现象，其背后几乎没有或完全没有经济现实依据。一家公司决定裁减员工，会受到其他公司也在裁员这一认知的影响。因此，我们研究裁员是否可以被建模为一种不受外部因素影响、自我维持的流行病。

日期：2023年5月10日。

1.4. 本文的结构如下。第 2 节介绍裁员数据集。第 3 节说明如何将这些数据建模为来自流行病学模型的一组含噪观测值。第 4 节说明如何将模型拟合到数据，并展示分析结果。第 5 节对这些结果进行分析。

2. Data

2.1. 我们从 Roger Lee 的网站 `layoffs.fyi` 获取了一组可自由使用的裁员数据。该数据涵盖了2020年3月11日至（截至数据提取时）2023年3月30日期间报告的2440起裁员事件。对于许多公司，数据中提供了被裁员工人数和/或公司裁员比例。然而，对于15%的公司，这两项统计数据均不可用。

2.2. 由于我们设想裁员潮会从一家公司传播到另一家公司，而不是从一名员工传播到另一名员工，因此我们选择关注发生裁员的公司数量，而不是被裁员工人数。每周裁员人数呈现出非常明显的每周模式。大多数裁员发生在星期二、星期三和星期四。因此，我们按周汇总裁员数据，以避免季节性问题。图1展示了每周数据。

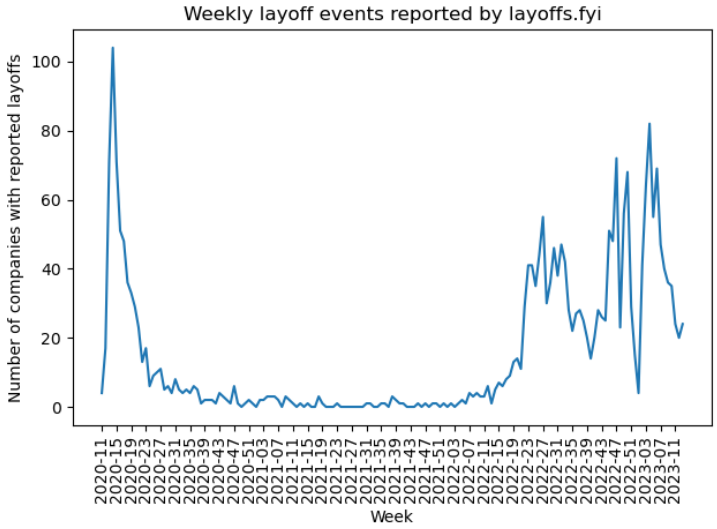


Figure 1. 每周裁员人数

2.3. 图1显示，在疫情期间达到峰值后，2021年全年的报告裁员人数都恢复到了非常低的水平。报告裁员人数于2022年再次开始攀升，至今尚未恢复到2021年的水平。

3. Model

3.1. 我们选择将裁员建模为始于2022年第1周的一场流行病。（另一个合理的选择可能是2022年第8周，对应于2022年2月24日俄罗斯入侵乌克兰。）

3.2. 我们假设有一个固定数量的 N 公司可能会裁减员工，并且相关裁员信息可能会被报道。这些 N 公司构成初始的易感人群。在起始时刻，部分易感人群处于感染期。处于感染期的公司是指将会裁员的公司。随着时间推移，越来越多的易感人群进入感染期。公司处于感染期一段时间后，会转入移除状态，且无法再回到之前的任一状态。因此，我们忽略多轮裁员的可能性，因为这通常只会发生在规模最大的公司中。

3.3. 我们假设，公司从易感状态转入感染期时便会发生裁员。因此，时间 t 的裁员人数等于时间 t感染期人群的瞬时增量。

3.4. 我们选择遵循Kermack和McEndrick[KM27]的SIR模型来描述模型的动态变化。这是最基本的流行病学模型之一。令 $S(t)$、 $I(t)$ 和 $R(t)$ 分别表示时间 t 时的易感者、感染期者和移除者群体。于是我们假设，对于某个 $\gamma, \delta > 0$，以下方程成立。

$$\begin{aligned}\frac{dS(t)}{dt} &= -\delta I(t)S(t)/N \\ \frac{dI(t)}{dt} &= \delta I(t)S(t)/N - \gamma I(t) \\ \frac{dR(t)}{dt} &= \gamma I(t)\end{aligned}$$

3.5. 在SIR模型下，时间 t 的裁员人数是 $I(t)$的瞬时增量。

$$\ell(t) = \delta I(t)S(t)/N \tag{3.1}$$

$j(t) = \ell(t)/N$ （在时间 t发生裁员的 N 公司所占比例）的近似解析表达式见 [KS20], 方程 33、42、54、58、59、60、61、62、66、71、73、80、81、82、83。它的中间部分呈高斯形状，两端则为指数型下尾和上尾。由于我们希望估算流行病的持续时间，因此尤其关注上尾。

$j(t)$ 的表达式仅取决于时间 0 时感染者所占的比例 $J(0)$，以及参数 $k = \gamma/\delta$。

3.6. 我们也可以将 $\ell(t)/N$ 建模为 $\int_{t-1}^{t} j(u)du$。这相当于观察第 $t-1$ 周至第 t 周之间的所有裁员，而不是在第 t 周进行一次快照。由于 $\int_{t-1}^{t} j(u)du \approx j(t)$，这种对 $\ell(t)$ 的替代选择对以下结果影响很小；如果将 $j(t)$ 替换为 $\int_{t-1}^{t} j(u)du$，第4.3节和第5.4节的最终结论也不会改变。

3.7. 假设时刻 t 观察到的裁员是时刻 t 实际裁员$\ell(t)$的随机样本。我们假设时刻 t 发生的每起裁员被报告的概率为 $p(t)$。因此，时刻 t 报告的裁员数服从二项分布 $\mathrm{Bin}(\ell(t)\ p(t))$。如果假设 $p(t)$ 与时间无关，那么模拟数据的方差将远低于观察数据的方差，因此我们必须允许 $p(t)$ 随时间变化。按照贝叶斯统计中的惯例，我们假设 $p(t)$ 是来自同一分布的独立抽样，并取该分布为 $\mathrm{Beta}(\alpha, \beta)$ 分布，其中 α 和 β 是模型参数。

3.8. 在这些假设下，如果 $x(t)_{t=0}^{T}$ 是观测到的裁员人数，那么 $x(t)$服从贝塔-二项分布，该模型的似然函数为

$$\prod_{t=0}^{T} \binom{Nj(t)}{x(t)} \frac{B(x(t)+\alpha, Nj(t)-x(t)+\beta)}{B(\alpha,\beta)}$$

其中 B是贝塔函数。该模型有五个参数： $J(0)$、 k、 N、 α、 β。参数$J(0)$和 k 决定了 $j(t)$。

3.9. 参数的最大似然估计值为 $J(0) = 0.00376$, $k = 0.899$, $N =$28261, $\alpha = 2.15$, $\beta = 2.92$。作为合理性检验，我们使用这些参数从模型中生成一些数据，并将其与观测数据进行比较。图2展示了一个示例。模拟的裁员人数（图2a中的虚线）具有随机性，因为它们取决于 $p(t)$。图2a中绘出了这些数据，以说明模型生成的数据在定性上与观测数据相似。图2b显示了一个范围，其中包含95%的模拟裁员人数。

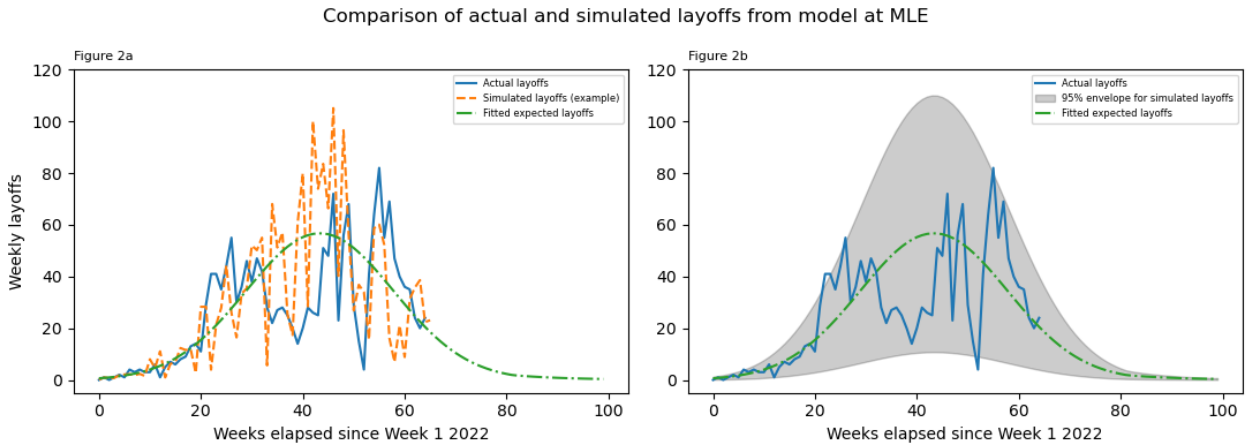


FIGURE 2. 数据与拟合模型的比较

4. RESULTS

4.1. 为了估算流行病持续的时间，我们需要确定流行病何时结束的定义。这个定义必须完全基于数据，因为我们所建模的是人们是否向 `layoffs.fyi`报告裁员，以及是否确实发生了裁员。一个自然的裁员基线水平是2021年每周报告裁员人数的平均值，即0.846。

4.2. 因此，我们将流行病结束时间定义为预期裁员人数首次达到 0.846 的时刻 t_{end}。时间 t 的预期裁员人数为 $Nj(t)\alpha/(\alpha+\beta)$，因此可以根据拟合模型的参数计算得出。

$$t_{\text{end}} = \min\{t : Nj(t)\alpha/(\alpha+\beta) < 0.846\}$$

4.3. 贝叶斯推断对每个参数采用均匀先验，具体如下：$J(0) \in [10^{-6}, 1]$; $k \in [0,1]$; $N \in [0, 10^{10}]$; $\alpha, \beta \in [0, 1000]$。运行 1000 次迭代的马尔可夫链蒙特卡罗得出了一个 95% 可信区间

$$t_{\text{end}} \in [96, 107]$$

这对应于裁员人数在2023年11月5日至2024年1月21日期间恢复至正常水平。结果如图3所示。

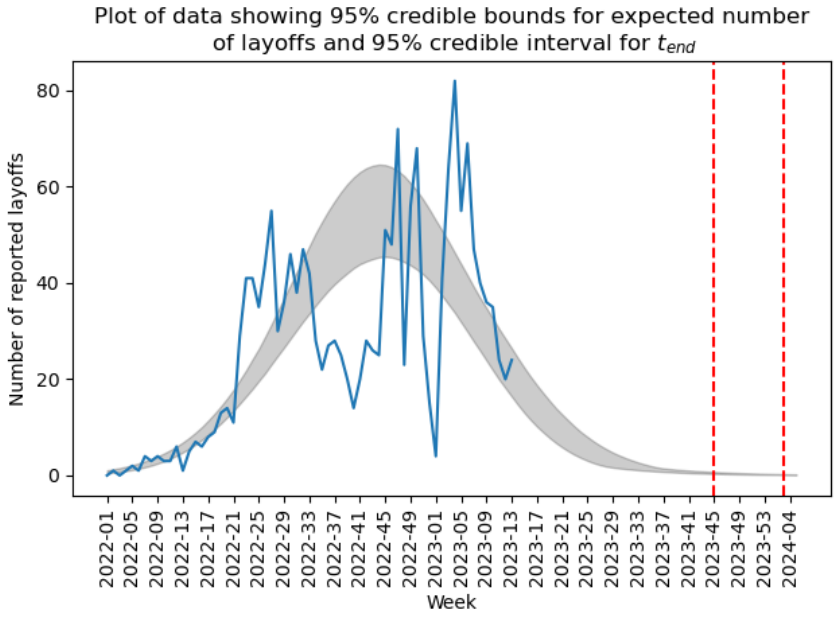


FIGURE 3. t_{end}的可信区间

5. DISCUSSION

5.1. 流行病模型的选择相当随意。该模型基于以下假设：企业之间均匀混合（这或许并非毫无道理），并且存在一个固定数量的公司会裁员，不会有新公司出现，也不会有公司倒闭或关停，且不会发生重复的裁员轮次。我们还选择按公司对裁员事件进行建模，而不考虑裁员的严重程度。

5.2. 此处提出的模型也忽略了数据中已知的非随机性来源。例如，2022年12月下旬裁员人数较少，可能是圣诞节和新年假期所致，但模型并未体现这一点。

5.3. 第4.1节中对 t_{end} 的定义选择也相当随意，并会影响结果。保守地说，我们预测科技行业裁员流行病将在2023年年底左右结束。

5.4. 为了检查数据对结果的影响程度，我们仅使用截至 t 的 $t \in \{50, 51, \ldots, 64\}$ 数据重复了分析。由此得出的可信区间绘于图4。看来，自2023年第3周以来，对 t_{end} 的预测一直相当稳定。

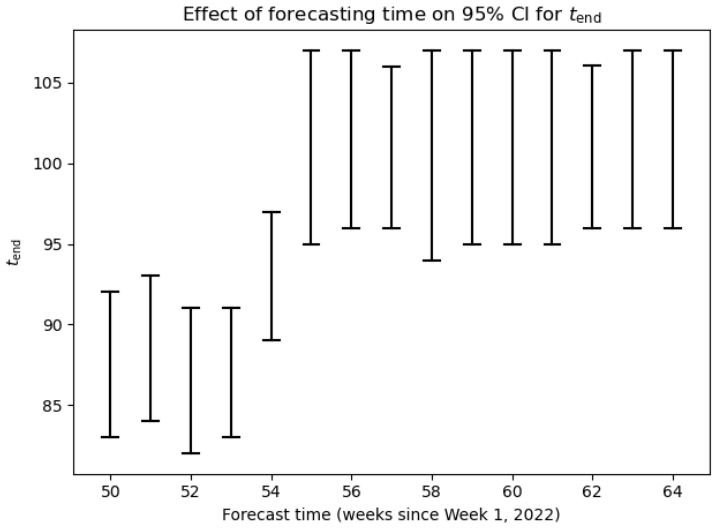


FIGURE 4. t_{end} 不同时点预测的可信区间

5.5. 完成本文后，我们又收集了四周的数据，以获得一个时段外样本。结果见图5。看来，对 t_{end} 的预测依然保持稳定，因此仍有理由预测，当前这波科技行业裁员流行病将在2023年底左右结束。这一预测是基于所选模型以及截至目前观测到的数据得出的最佳估计。

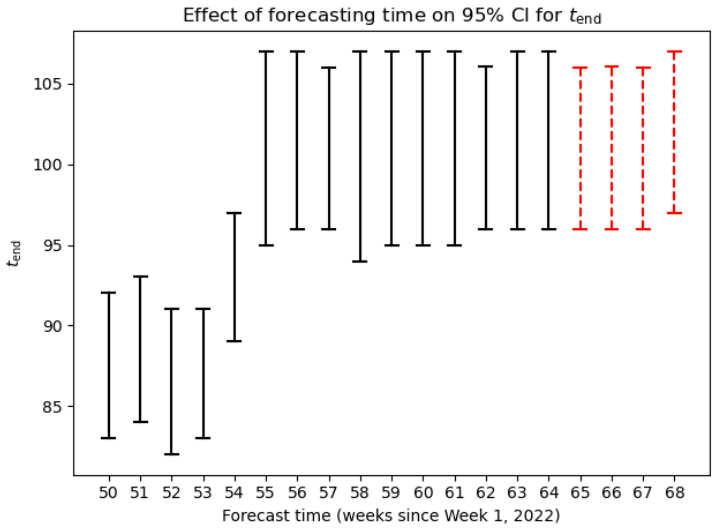


FIGURE 5. t_{end} 不同时点预测的可信区间，包括额外四周的数据

6. Conclusion

6.1. 自2022年初以来，由一款在线裁员追踪器报告的裁员情况，似乎遵循了SIR模型中传染病在科技公司之间传播时产生的噪声观测值模式。这提示裁员可能既受到社会因素的推动，也受到经济因素的推动。对裁员情况进行模型拟合表明，目前的高裁员期将在2023年年底左右结束。到2023年年底，我们将能够检验这一预测是否正确。

7. Acknowledgements

7.1. 感谢Horace Choi博士提出宝贵意见。论文 [Rod16] 列举了SIR模型在流行病学之外的一些其他应用。

References

[dW22] Melissa de Witte。为什么会有这么多科技行业裁员？我们为什么应该对此感到担忧？斯坦福学者对此进行了解释。斯坦福新闻，2022年。[HML+23] Md Nayem Hossen、Md Alamgir Mollah、Nusrat Sharmin Lipy、Gazi Md Shakhawat Hossain和Md Shahinur Rahman。高科技产业裁员的影响因素：来自美国的证据。《国际信息、商业与管理杂志》；中坜，15(1)，2023年。[KM27] William O Kermack和Anderson G McKendrick。对流行病数学理论的贡献。《伦敦皇家学会会刊*A*辑：数学、物理与工程科学》，115(772):700–721，1927年。[KS20] M Kr¨oger和R Schlickeiser。SIR模型在流行病时间演化中的解析解。A部分：与时间无关的再生因子。《物理学杂志*A*：数学与理论》，53(50)，2020年。[OEC14] 经济合作与发展组织。金融危机期间及之后的经济合作与发展组织预测：事后分析。《经济合作与发展组织经济部政策简报》，(23)，2014年。[Rod16] H.S. Rodrigues。SIR流行病学模型的应用：新趋势。《应用数学与信息学国际杂志》，10:92–97，2016年。[SW23] Sandra J. Sucher和Marilyn Morgan Westner。科技行业裁员：2022年–2023年。《哈佛商学院案例 *323-095*》，2023年。

电子邮箱地址：richardvale@d24h.hk

Laboratory of Data Discovery for Health, Units 1201-1206, 1223 & 1225, 12/F, Building 19W, 19 Science Park West Avenue, Hong Kong Science Park, Pak Shek Kok, New Territories, Hong Kong.