使用动态对照匹配验证大规模并行自适应测试

Schaun Wheeler

Aampe 公司

美国北卡罗来纳州罗利

schaun@aampe.com

摘要

A/B 测试是营销优化中广泛使用的一种范式，因为它能够识别因果效应，而且大多数消息投递软件平台都提供开箱即用的实现。然而，现代企业往往同时并行开展许多 A/B/n 测试，并在同一条消息中打包多个内容变体，其中并非所有变体都属于明确的测试。无论这是因为多个团队同时进行测试，还是源于更复杂的强化学习（RL）方法——该方法会根据先前结果持续调整测试和测试条件分配——动态并行测试都无法像传统 A/B 测试那样进行评估。本文提出一种方法，利用与测试同步调整的匹配合成对照组，在持续测试适应的条件下分离各种测试的因果效应。

关键词： A/B 测试、多臂老虎机算法、强化学习、合成控制、粗化精确匹配

1 引言

企业很少能够对所有可用的机会进行投资。企业可能没有足够的资源投资于所有选项；投资某一选项可能会排除对其他替代选项的投资；

而负责行使这些选择权的人员可能只是没有足够的时间或精力来处理所有事情。由于这些限制，企业通常会尝试从所有可用选项中选出一个子集，以最大化其投资回报。能否进行优化的子集选择，可以说是决定企业经营状况以及增长速度的一个主要因素。

营销沟通受到诸多限制，因此有必要进行优化的子集选择。（如果一家企业始终在任何时间向所有客户传达所有信息，那么它将失去所有客户，同时还会因此产生无法管理的消息发送成本。）然而，选择最优的消息发送参数——时间安排、主题、文案细节等——会给营销决策者的决策能力带来巨大压力。识别可用选项、为每个选项设置评估、分析结果，并将这些结果转化为后续运营决策（而这些决策本身也应接受评估），都需要投入大量人力。

大多数用于电子邮件、短信、推送通知及其他渠道的营销传播交付平台都具备开展 A/B测试的功能。A/B测试的基本原理和流程很简单：（1）选择两个选项，作为实验中的处理方案；（2）将一组联系人（客户、应用程序用户等）随机分配到每种处理方案中；（3）实施处理方案，换句话说，就是向联系人发送消息；（4）衡量打开率、点击率、加入购物车率、购买率、留存率或任何其他对业务具有重要意义的指标。成功指标表现更好的处理方案就是更优的选择。

关于实验者需要设定多大的差距，才能认定胜出方案的效果确实优于落败方案，存在许多不同的“最佳实践”。然而，这些验证方法往往为时过早，因为最常见的 A/B 测试实施方式未能考虑各种抽样偏差和测量工具偏差。其中最突出的问题包括：（1）高价值联系人仅仅由于随机偶然因素而被不均衡地分配到某一种特定处理组 [1, 2], ；（2）未能考虑到以人类可读消息的形式传递处理时必然引入的处理差异 [3]。此外，即便测试结果稳健，也可能因季节性、用户适应或其他影响而仅具有短期价值 [4, 5]。简而言之，A/B 测试很容易让企业扭曲或误读测试结果。

多种源自强化学习（RL）的方法——尤其是多臂老虎机和上下文老虎机——能够解决 A/B测试固有的弱点，同时保留其有用的推断特性。强化学习方法可以监测所有可能影响联系人行为的消息发送选项，并跟踪后续行为，从而自动为联系人群体的不同子集选择影响力大的选项。由于强化学习方法本质上具有自适应性，而且能够同时考虑许多不同的潜在影响因素，因此可以缓解以下问题：

处理分配不对称、处理的混杂和/或调节作用，以及测试结果的持久性。

任何持续自适应系统面临的一项挑战都是验证，因为这类系统没有可供比较结果的“之前”和“之后”阶段。为了解决这一难题，我们提出了一种构建自适应控制组的方法。首先，我们简要概述一种在实践中实现强化学习系统的方式，使其能够在应用程序用户生态系统中自适应地优化消息发送，从而产生系统性影响。随后，我们评估一些传统上用于衡量系统性影响的其他方法，并说明自适应控制如何克服这些方法的不足。接着，我们介绍构建控制组的指导原则，并展示一家企业用户使用真实世界强化学习消息发送系统所取得的结果。

2. 实现强化学习消息发送流程的一种方法

下面简要概述我们如何采用强化学习而非 A/B测试来实现持续自适应的消息发送流程。[1] 我们已在别处充分论证了这一学习架构 [6]。这只是一个高层次的概述，当然也只是构建此类系统的一种可行方式。

2.1 一般行为模式的表示。

该流程首先必须具备监测用户行为的能力。这意味着系统可以部署在应用程序或网站的事件流之上，仅使用第一方数据。对于电子商务应用或网站，记录的事件可能包括登录、更新个人资料、查看商品、将商品加入购物车、开始结账流程、完成购买等活动，也可能包括更底层的事件，例如点击特定按钮或查看特定推荐内容。

第一方数据最重要的特征之一是，这些事件不仅都带有时间戳，还带有一个持久 ID，用于标识唯一的联系人。这意味着可以针对不同时间范围（如一周、一个月等），为每位联系人计算事件摘要，从而概括每位联系人的“正常”行为模式。如有必要或需要，可以采用主成分分析等降维技术，为每位联系人的活动生成精简表示，从而降低后续分析的计算复杂度。

2.2 交织式处理分配

[1] https://aampe.com

基于强化学习的学习系统无需识别仅包含少数几种处理方式的离散“测试”。[2]在现实世界的消息发送场景中，这类测试无论如何都不可行，因为必须将处理方式整合为人类可读的文字内容，发送给联系人。例如，一条消息可能包含强调便利性价值主张，并采用简洁、商务化的语气；另一条消息可能包含相同的价值主张，但采用轻松、鼓励性的语气；还有一条消息可能包含强调低成本的价值主张，但采用商务化的语气；等等。手动撰写消息以涵盖所有可能的处理方式组合并非易事。不过，可以创建多个符合各项预期处理方式的文本片段，再将这些片段组合起来，确保可以向联系人发送包含不同处理方式组合的消息。[3]

当系统对某项处理的预期效果有较高置信度的估计时，就会更频繁地发送相应消息。[7] 如果不存在明显效果更好的估计，系统就会选择一条能更充分探索处理空间的消息。无论是探索新的可能性还是利用过去的经验，对决策过程进行随机扰动都能避免优化陷入局部最大值。

2.3 行为监测与奖励定义

与标准 A/B 测试类似，强化学习系统可以使用预先设定的监测窗口（例如，消息发送后 24 小时），以确定哪些联系人在应用程序中执行了哪些操作。根据需要，不同事件可为系统带来不同的奖励值，也可以根据消息与事件之间的时间间隔对奖励进行折扣。例如，假设系统最希望联系人完成购买，那么在一小时内完成购买可能会获得较高的奖励权重；一小时内访问应用程序但未完成购买可能会获得较低的权重；消息发送八小时后完成购买也可能获得较低的权重；八小时后访问应用程序但未完成购买，则可能获得最低的权重。

2.4 基于模型的推断

我们在此讨论的系统会针对每位联系人进行优化——本质上，这是一个情境赌博机，其中每位用户都是一个独立的情境。因此，推断是在唯一用户 ID 的层面上进行的。在更传统的情境赌博机中，即使某个情境代表了总人口中相对较大的一个群体，也可以针对每种情境进行推断。不过，为了让下文的讨论更加简单清晰，我们假设情境就是单个用户。

---

[2] 根据我们的经验，实施多臂老虎机的公司通常会继续将不同消息视为离散的处理，这或许是因为这样更容易进行思考和评估绩效。我们认为，这会使大量价值无法实现，因为相互交织的处理能够支持推断，从而实现交叉销售，以及将一条消息的结果应用于其他消息等功能。[3] 例如，请参阅：
https://www.aampe.com/blog/how-to-get-more-engagement-from-your-existing-customers-without-making-any-new-content

对上述动态系统进行建模的一种简单方法，是创建一个训练集，其中包含历史行为（或其潜在空间表示）、用户接受过的处理（即使同一条消息中包含多重处理），以及模型先前对每位联系人的处理偏好的估计。如果已计算奖励权重，则可以据此对训练数据加权。随后，可以将其建模为一个分类问题：消息促成目标事件时，将其归为正类；未能促成目标事件时，则将其标记为负类。如果使用了奖励权重，还可以通过设定阈值，将这些权重划分为正类和负类。

使用模型的好处不仅在于它能够纳入比人工操作员合理处理范围更广泛的信息，还在于它能够估算两种不同的预测：（1）仅根据历史特征得出的目标达成基准概率；（2）针对每种可能的处理，估算其边际概率，并将该处理及联系人对该处理的偏好与历史行为一并纳入模型。随后，将目标估算值与基准估算值进行比较，以估算联系人在接受某项特定处理后达成目标的概率是否高于基准预期。[4]

2.5 前馈分配

这些用户偏好估算值可以存储在系统中，用于后续的处理分配和评估。学习到的偏好可用于选择时间（星期几、一天中的时段等）、主题（产品类别、品牌等）、文案（价值主张、行动号召、语气等）和渠道（推送通知、电子邮件、短信、应用程序内通知等），也可用于操作员可能希望改变消息发送方式的其他方面。

3. 留出组与系统性影响

上一节概述的系统是一种持续自适应的替代方案，可替代有时间限制的 A/B测试。因此，评估此类系统比评估简单的 A 与 B 选项更具挑战性，也比评估政策变更等一次性干预的影响更具挑战性。本节将回顾三种估算系统性影响的方法，并讨论为何这些方法都不适合评估动态消息发送系统。

全局留出组。 最简单的对照组形式是留出样本。在这种方法中，一部分联系人将完全不参与消息发送。要创建一个与其他记录保持可比性的留出组并不容易 [8]。此外，留出组规模越大——而它必须足够大，才能

[4] 有许多不同的方法可以合理地构建模型，以估算用户参与特定奖励行为的倾向是否高于基线。评估这些不同的建模选择超出了本文档的范围。有关一组可能选择的更多信息，请参阅 https://www.aampe.com/blog/how-do-i-know-it-works-technical-explanation

在分析上发挥作用——实验干预所能带来的收益就越少 [9]。全局留出组不适用于评估动态系统。在营销消息发送的情境中，如果经过妥善筛选和维护，全局留出组可以帮助得出结论，了解消息发送相较于从未接收过消息的益处。但这并不是现实的业务场景。重要的问题是，消息发送能否在当前现状的基础上带来改善。在自适应迭代分析中，全局留出组可能导致对留出组过拟合 [10]。

交替测试。 交替测试会交替安排联系人接受处理的时间段[11, 12, 13]。例如，某一天一半联系人会收到一条消息，另一半则什么也收不到。第二天，之前没有收到消息的那一半会收到前一半收到的消息，而前一半则什么也收不到。交替测试是控制网络效应的有效方法，但无法检测学习效应。在自适应系统中，学习效应是我们预期会出现的主要效应之一 [14]。

合成控制。 合成控制用于识别不会受到处理本身影响的结果预测因子。根据这些因素拟合模型，以预测结果 [15]。随后实施处理，将实际结果与预测结果之间的差异作为因果效应的估计值。合成控制是一种强大的工具，但有两个明显的缺点。首先，因果效应的估计取决于能否识别独立预测因子，以及能否根据这些预测因子准确地建立结果模型。这类预测因子未必总是可用，而且模型拟合优度也总是存在争议 [16]。另一个问题是，合成控制与全局留出组类似，能够模拟从某个初始且未变化状态开始的变化，但在不断变化的系统中，很难模拟从近期用户状态开始的变化。合成控制更适用于一次性干预。

所有用于构建控制组以估计干预因果影响的方法都各有明确优势，但在应对本文所描述的动态自适应系统方面仍有所欠缺。

4. 通过粗化精确匹配进行自适应控制

我们构建控制组的方法借鉴了标准留出实践中的三项原则：(1) 控制组不得接触过处理，(2) 可以使用时间窗口限定接触范围，以及 (3) 可以根据控制案例与处理案例之间的相似性明确选择控制案例，以确保两者具有可比性。

粗化精确匹配（CEM）是一种用于因果推断研究的统计方法，旨在减少非随机抽样造成的抽样偏差 [17]。该方法以分层随机抽样为理论基础

随机抽样 [18], CEM 在不同研究人群中相似个体组成的群体之间建立近似匹配。接受处理的个体会根据这些分层，与未接受处理的用户进行匹配。

我们使用粗化精确匹配来创建一个会随着测试人群变化而调整的合成对照组。从理论上讲，任何因素组合都可以用于匹配。我们已成功实现了一种会考虑以下因素的对照方法：

1. **时间窗口**。 每条发送的消息都会与一位用户进行匹配，该用户 (1) 在相关消息发送前 2-8 天收到过一条消息，且 (2) 在相关消息发送前 1 天至发送后 1 天之间没有收到任何消息。第一项标准通过匹配近期发送的消息来减少季节性效应，第二项标准则减少因同时实施并评估多重处理而产生的测量工具偏差。

2. **消息簇**。 可以根据词语相似度、消息中包含的处理，或其他任何特征对消息进行聚类。这样做的目的是捕捉消息之间的某种相似性，从而使被分配到某条特定消息的联系人能够与近期收到过同类型消息的联系人相匹配。

3. **用户活跃度**。 可以根据任意数量的特征对应用程序或网站用户进行分类。在实践中，我们发现，根据以下两个方面对用户进行分类很有用：（1）用户距离上一次发生“转化”事件有多久——在电子商务中，这指的是购买；（2）用户距离上一次访问商店或应用程序有多久。这两种活动分别分为“活跃用户”行为（过去 7 天内发生）、近期用户行为（过去 8–30 天内发生）和“潜伏”行为（过去 31–90 天内发生）。此外，还可以设置一个新用户类别（过去 7 天内首次出现在应用程序或网站上的用户）。[5] 根据这些类别进行匹配，可确保用户只会与具有相似基线行为水平的其他用户进行匹配。

一旦将控制用户与一条消息匹配，就可以监控这两个联系人的应用程序和网站行为（不过实际上只有测试消息的联系人会收到任何通信）。

[5] 此处列出的天数阈值仅为示意。可以使用任意阈值。

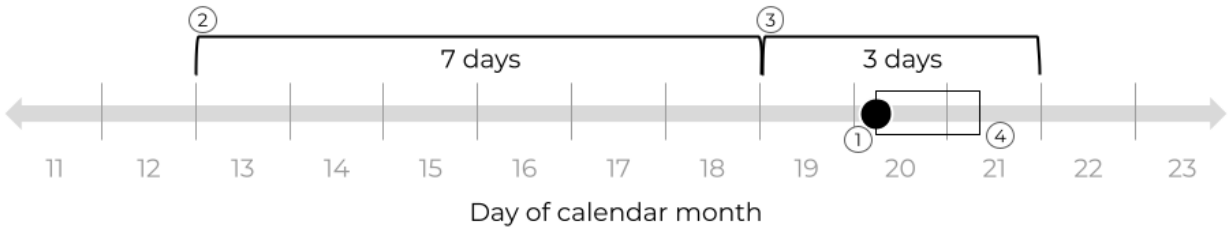


图 1：时间窗口示例。默认情况下，所有对照候选项均以一条测试消息为参照进行选择 (1)。如果对照候选项在测试消息发送前一天之前的 7 天内收到过消息 (2)，且在测试消息发送前一天、发送当天及发送后一天均未收到消息，则会被选中 (3)。测试消息及匹配对照的效果会在测试消息发送后的 24 小时内进行监测 (4)。

粗化精确匹配方法的优势在于，除了匹配后的测试组和对照组具有良好的统计特性外，它不需要建立模型，因此研究结果也无需根据拟合优度指标进行调整。当分层中样本不足以进行匹配时，粗化精确匹配的结果并不可靠；不过，可以轻松检查分层内容并进行调整，以确保存在可匹配的样本。

5. 结果案例研究

以下结果来自一款大型电子商务应用。我们采用了默认的用户活动类别定义：

1. 新用户：使用应用程序或网站不超过七天。2. 活跃用户：距上次转化或访问事件不超过七天。3. 近期用户：距上次转化或访问事件八至三十天。4. 曾活跃用户：距上次转化或访问事件三十一至九十天

由于粗化精确匹配不依赖模型，因此无需考虑拟合优度指标。不过，希望获得一些匹配适当性的保证是合理的。表 1 展示了评估匹配质量的两种不同方法。

表 1：匹配质量指标

		Messages			Contacts		
Category	Subcategory	Test	Control	Coverage	Test	Control	Coverage
New	New	259,076	234,633	90.6%	36,125	2,208	6.1%
Active	Actively converting	242,983	242,949	100.0%	61,782	28,856	46.7%
Recent	Recently converted	393,930	393,924	100.0%	98,212	70,677	72.0%
Lurking	Formerly converted, actively looking	234,056	234,055	100.0%	85,194	52,682	61.8%
Lurking	Formerly converted, recently looked	54,642	54,642	100.0%	35,844	28,282	78.9%
Lurking	Actively looking	1,698,175	1,698,175	100.0%	443,722	296,271	66.8%
Lurking	Recently looked	791,884	791,884	100.0%	375,422	312,630	83.3%
Inactive	Formerly converted, formerly looked	10,857	10,852	100.0%	7,194	6,202	86.2%
Inactive	Formerly looked	193,846	193,846	100.0%	109,178	92,568	84.8%

“测试”消息是实际发送给联系人的消息。对于每条此类消息，我们都会采用上一节所述的方法，搜索一份可行的对照用户列表。我们会随机选择其中一名对照用户，系统会为该用户记录一条“对照”消息，并记载为与其匹配的测试消息同时发送，但实际上这条消息根本没有发送出去。对于至少找到一名对照用户的测试消息，其所占百分比是衡量粗化精确匹配流程质量的一项指标。在大多数情况下，这一百分比接近100%。请注意，对于新用户而言，这一比例会稍低一些，因为新用户的时间窗口非常小。

评估粗化精确匹配流程质量的另一种方法，是计算对照联系人与测试联系人的比率。在任意一段时间内（上表采用的是14天），同一个联系人可能会收到多次消息——在我们此处评估的系统设置下，一周内最多可能收到35次，但通常更接近每周2或3次。对于这个具体案例，新联系人的对照联系人与测试联系人之比相当低。这意味着同一个对照联系人会被随机匹配到许多不同的测试联系人，这表明新用户的对照联系人池相当小。查看原始数字可以发现，新联系人仍有超过2,000名测试联系人。对于计算汇总统计量而言，这是一个相当大的样本量。不过，对于这一具体案例的新用户，测试组与对照组的比较可能仍应比其他群体的比较更加谨慎地看待。

由于该系统运行的是持续演进的实验，而非离散且有固定时限的 A/B测试，因此我们无法在某个单一时点简单地收集“结果”，以展示测试记录相对于对照记录的表现。在目前讨论的系统中，消息决策是基于决策置信度分数作出的，这些分数概括了各个联系人对不同消息发送时间或文案决策的偏好。换句话说，每项消息决策都是在检验这样一个假设：系统已准确估计出每位联系人的消息偏好。

我们可以根据消息的置信度评分，将消息分入规模适当的样本组，从而利用这些评分动态评估相对表现。具体而言，最低分的消息进入一组，次低分的进入下一组，依此类推，直到最高分的消息进入最高分组。每组的大小可以灵活设定——我们通常每组包含 1000 条消息。[6] 消息分组后，我们便可以计算任意表现指标，并查看测试组与对照组的表现估算[7],，如图2所示。

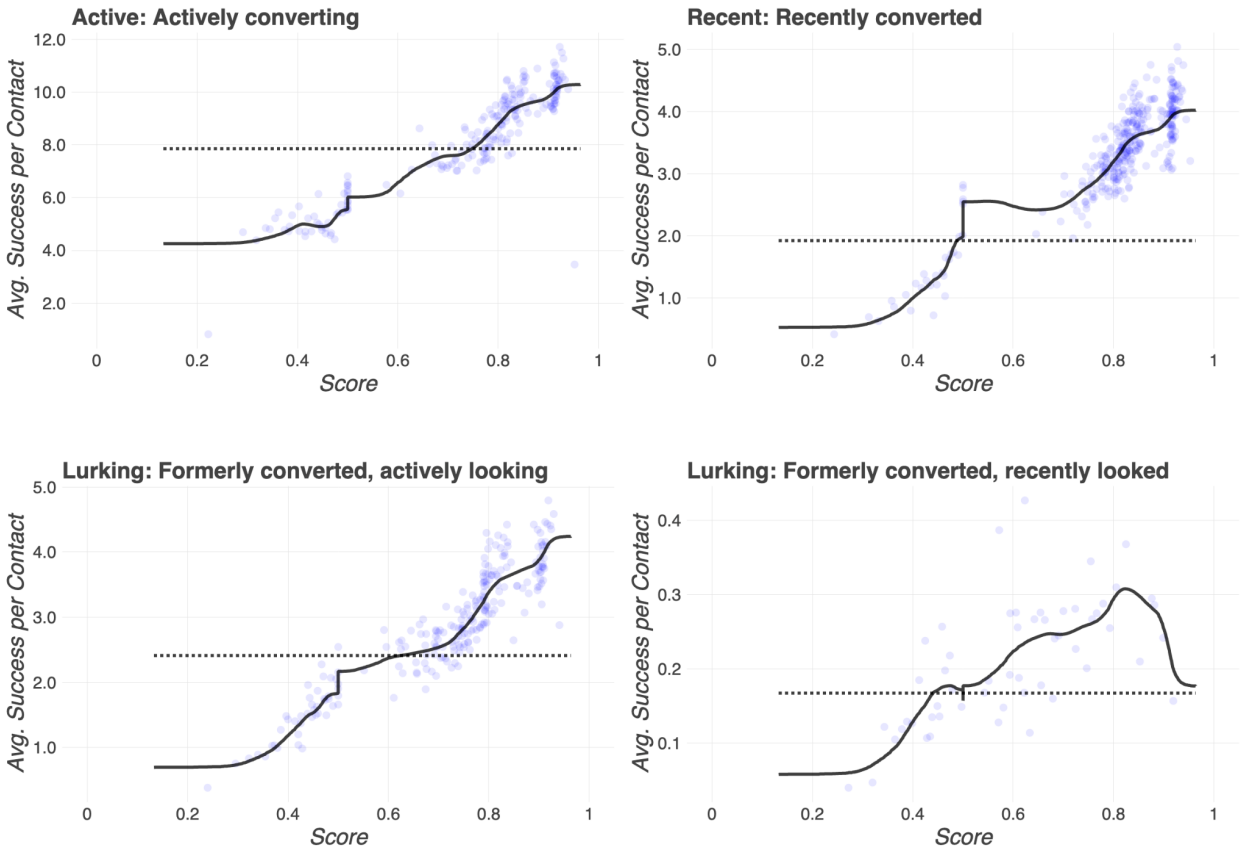


图2：四种不同活动类别中，每个分箱内每位联系人的应用程序/网站访问事件平均次数。虚线表示对照值。蓝点表示测试值。实线表示平滑后的测试值。

如图所示，对照指标（虚线）在各个区间内保持不变。这正是我们对照组所预期的情况，因为对照联系人是随机匹配到测试消息的。总体而言，置信度得分越高，测试消息的表现就越优于对照组。

[6] 无法保证待分箱的消息数量是 1000 的倍数。如果某个分箱中的数量少于 1000，我们在计算成功指标时，会将实际数量与 1000 之间的差值视为先验。实际数量与 1000 相差越大，我们就越会将指标收缩至 0。因此，如果某项成功指标基于一个仅包含 100 个样本的分箱得出，数值为 100%，那么修正后的指标将为 (100% * 0.1) + (0% * 0.9) = 10%。[7] 我们通常将联系人收到消息后 24 小时内触发的应用程序或网站事件平均次数作为成功指标。默认跟踪的事件包括应用程序访问（访问应用程序或网站的任何行为）、加入购物车（表明购买意向的强烈信号，但并非转化事件）以及产生收入（实际的转化事件）。这些名称是针对电子商务用例调整的，而非电子商务用例通常也有类似的事件需要监测。

消息。这正是我们所预期的：置信度评分启发式方法确实捕捉到了有关消息决策的信息，而这些决策会提高联系人回复的可能性。

对于图2所示的四个类别中的三个，测试消息从约 0.5 分处开始优于对照消息。这也在意料之中，因为 0.5 是置信度评分的初始值，表示联系人对特定的消息发送决策既没有正向偏好，也没有负向偏好，因此他们响应此类决策的概率为 50%。然而，对于“活跃用户”（即在消息发送前 7 天内完成过转化事件的用户），测试消息要到置信度评分更高的区间——接近 0.75 分时——才开始优于对照消息。这是因为活跃用户已经处于执行目标事件的状态。他们的基准水平更高，因此需要影响力更强的消息，才能使联系人表现超出这一已经很高的基准水平。

图3展示了相同的结果，但我们不再评估消息发送对应用程序或网站访问量的影响，而是展示其对创收事件（已完成结账）的影响。

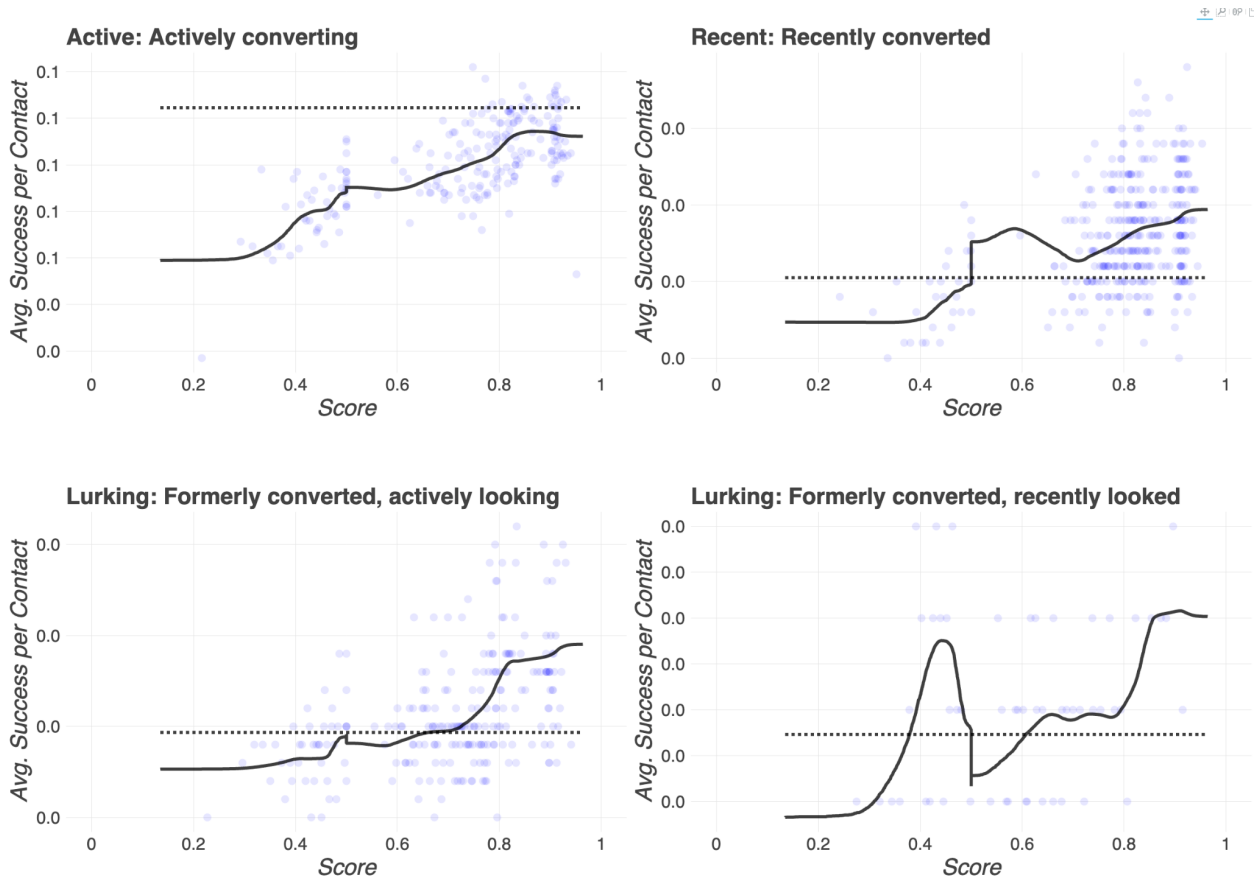


图3. 四种不同活动类别中，每个分箱内每位联系人的结账完成事件平均数量。虚线表示对照值。蓝色圆点表示测试值。实线表示平滑处理后的测试值。

这些结果表明，或许并不令人意外，成功指标会因成功的定义方式不同而大相径庭，表现也各不相同。只要某种联系人行为本身不依赖于消息发送，就可以使用动态对照组来评估消息发送对该行为的影响。例如，动态对照组无法衡量对点击率的影响，因为按照定义，没有收到任何消息的联系人不可能点击消息。

6. 关于归因的说明

几乎所有企业都关心营销或广告活动的归因。动态对照组可以直接用于估算归因。图2和图3中每个分数区间均代表1000条消息，同时也包含发送这些消息后24小时内观察到的成功事件数量（应用或网站访问、加入购物车、已完成购买等）。如果某个区间的对照组成功指标高于测试组指标（在图中，即虚线高于某个蓝点），那么显然不能将该成功归功于强化学习算法。不过，如果某项成功指标高于对照组指标，那么测试组超出对照组的部分就可以归因于消息发送。

例如，如果某个分组中测试消息带来的应用程序访问次数平均为 4.3 次，而对照消息带来的应用程序访问次数平均为 1.7 次，那么该分组中测试消息带来的成功事件中，有 (4.3 - 1.7) / 4.3 = 60% 可归因于这些消息，而非随机因素或基线预期。表2显示了各用户活动类别中可归因的成功事件百分比。

表2：可归因于消息发送的成功事件百分比

Category	Subcategory	App visited	Item added to cart	Revenue Generated
New	New	5.6%	23.5%	1.0%
Active	Actively converting	10.6%	6.5%	0.3%
Recent	Recently converted	40.9%	37.6%	29.4%
Lurking	Formerly converted, actively looking	18.6%	17.5%	19.8%
Lurking	Formerly converted, recently looked	21.7%	25.6%	31.2%
Lurking	Actively looking	11.8%	12.1%	20.0%
Lurking	Recently looked	11.8%	15.1%	15.9%
Inactive	Formerly converted, formerly looked	0.0%	7.9%	12.0%
Inactive	Formerly looked	6.0%	9.7%	7.6%

如表所示，在此用例中，消息发送对近期完成转化、但在消息发送时并未积极转化的用户影响最大。消息发送也产生了相当大的

对较早完成转化、但近期访问过网站或应用程序的用户产生了影响。不同成功指标所受的影响也各不相同。消息发送对促使已经活跃的用户访问应用程序的作用，远大于对促使他们购买的作用；但对于很久以前购买过、近期访问过应用程序的用户，消息发送对购买行为的影响则大于对访问行为的影响。

将各区间中可归因百分比乘以成功事件的实际数量，再对各区间求和，我们便可估算出归因于消息发送的增量访问次数、加入购物车次数或购买事件数。所有这些都取决于监测窗口的选择，只有在该窗口内发生的事件才会被视为归因候选。

7. 结论

在实际应用的几乎所有方面，A/B 测试都逊于基于强化学习的动态干预决策。然而，A/B 测试不仅无处不在，而且很容易得出易于解读的结果。正如本文开篇所讨论的，这些结果往往是虚假的，或被夸大了；但对于非技术人员来说，容易误解、淡化这一事实，甚至刻意忽视它，因为他们需要向同事、经理和雇主证明日常决策的合理性。过去，动态调整的实验很难评估，除非采用全局留出组并监测长期效果（本质上是把整个动态系统变成一次规模巨大的 A/B 测试）。本文展示的基于粗化精确匹配的动态控制分配表明，我们可以监测这些系统的即时效果，并据此得出有意义的归因估计。

参考文献

[1] Wang, Yu and Gupta, Somit and Lu, Jiannan and Mahmoudzadeh, Ali and Liu, Sophia. A/B 测试中的重度用户偏差。*arXiv* 预印本 *arXiv: 10.48550/ARXIV.1902.02021，2019*。

[2] Davidson-Pilon, Cameron. A/B 测试中的类别不平衡问题。*https://dataorigami.net/2015/11/22/The-Class-Imbalance-Problem-in-A-B-Testing.html*

[3] Wang, Jingshu andZhao, Qingyuan and Hastie, Trevor and Owen, Art B. 多重假设检验中的混杂因素调整。*arXiv* 预印本 *arXiv: 10.48550/ARXIV.1508.04178，2015*年。

[4] Deng, Alex、Li, Yicheng、Lu, Jiannan 和 Ramamurthy, Vivek。《A/B 测试中的选择后推断》。载于第27届 ACM 知识发现会议论文集，arxiv: 10.1145/3447548.3467129，2021年。

[5] Sadeghi, Soheil 和 Gupta, Somit 和 Gramatovici, Stefan 和 Lu, Jiannan 和 Ai, Hao 和 Zhang, Ruhan. 新颖性与首因效应：在线实验的长期估计器。arxiv 预印本 arxiv: 10.48550/ARXIV.2102.12893，2021年。

[6] Wheeler, Schaun。《我如何知道它有效？》Aampe 学习系统的（某种程度上）技术性解释。访问日期：2023年2月8日，网址：https://www.aampe.com/blog/how-do-i-know-it-works-technical-explanation

[7]Surmenok，Pavel。《上下文老虎机与强化学习》。访问日期：2023年2月24日，网址：https://towardsdatascience.com/contextual-bandits-and-reinforcement-learning-6bdfeaece72a

[8] Bennett, Michele 和 Nekouei, Mehdi、Mehta, Armand、Prieditis Rajesh、Kleczyk, Ewa 和 Hayes, Karin。创建用于医疗保健领域机器学习分析的分析朴素留出记录以及训练和测试记录的方法。*arXiv* 预印本 *arXiv: 10.48550/ARXIV.2205.03987, 2022*.

[9] Haidar-Wehbe、Sami、Emerson、Samuel R、Aslett、Louis J Mand Liley、James。用于安全预测模型更新的留出集最优规模。arXiv预印本 arXiv: 10.48550/ARXIV.2202.06374，2022年。

[10] Dwork、Cynthia、Feldman、Vitaly、Hardt、Moritz、Pitassi、Toniann、Reingold、Omer 和 Roth、Aaron。《自适应数据分析与留出集复用中的泛化》。arXiv 预印本，arXiv: 10.48550/ARXIV.1506.02629，2015年。

[11] Bojinov、Iavor 等人。《切换实验的设计与分析》。SSRN电子期刊，2020年。Crossref, https://doi.org/10.2139/ssrn.3684168。

[12] Yuchen Hu 和Stefan Wager。《几何混合下的切换实验》。arXiv 预印本 arXiv: 2209.00197, 2022年。

[13] Li、Hannah 等人。“双边市场实验中的干扰、偏差和方差：平台指南。”《2022 年 ACM Web 会议论文集》，2022 年 4 月。Crossref，https://doi.org/10.1145/3485447.3512063。

[14] Feng、Wei 和Bauman、Jared。《实验中平衡网络效应、学习效应与统计功效》。访问日期：2023年2月9日，网址：https://doordash.engineering/2022/02/16/balancing-net work-effects-learning-effects-and-power-in-exp eriments/.

[15] Abadie, Alberto 和 Vives-i-Bastida, Jaume。《合成控制的应用》。arXiv 预印本 arXiv: 2203.06279，2022年。

[16] Martinez, Ignacio 和 Vives-i-Bastida, Jaume。合成控制的贝叶斯推断与频率学派推断。arXiv 预印本 arXiv: 2206.01779，2022年。

[17] Iacus, Stefano M. 和 King, Gar 和 Porro, Giuseppe。2012。“无需平衡性检验的因果推断：粗化精确匹配。”*Political Analysis*，20，1，第 1–24 页。

[18] Iacus, Stefano M.、King, Gar 和 Porro, Giuseppe。2019。“因果研究中匹配方法的统计推断理论。”*Political Analysis*，27，1，第 46--68 页。

[19] Ripollone, John E 和 Huybrechts, Krista F 和Rothman, Kenneth J 和 Ferguson Ryan E 和 Franklin, Jessica M. 2020年。“使用真实和模拟理赔数据评估粗化精确匹配在药物流行病学中的效用”，《美国流行病学杂志》，第 189 卷，第 6 期，6 月，第 613–622 页。