图神经聚类

ABSTRACT

对敲交易作为一种常见的交易量操纵形式，旨在吸引投资者并误导其做出错误投资判断，在ERC20加密货币中尤为突出。为此，本文提出两种基于ERC20加密货币链上交易数据的算法，以保留对敲交易的直接证据。在对对敲交易进行标记后，进一步获取其特征并量化交易规模。实验表明，大多数ERC20加密货币的对敲交易率超过15%，其中UNI代币超30%的交易被标记为对敲交易。这说明多数ERC20加密货币的活跃度不真实，恢复真实数据对市场监管至关重要。

Introduction

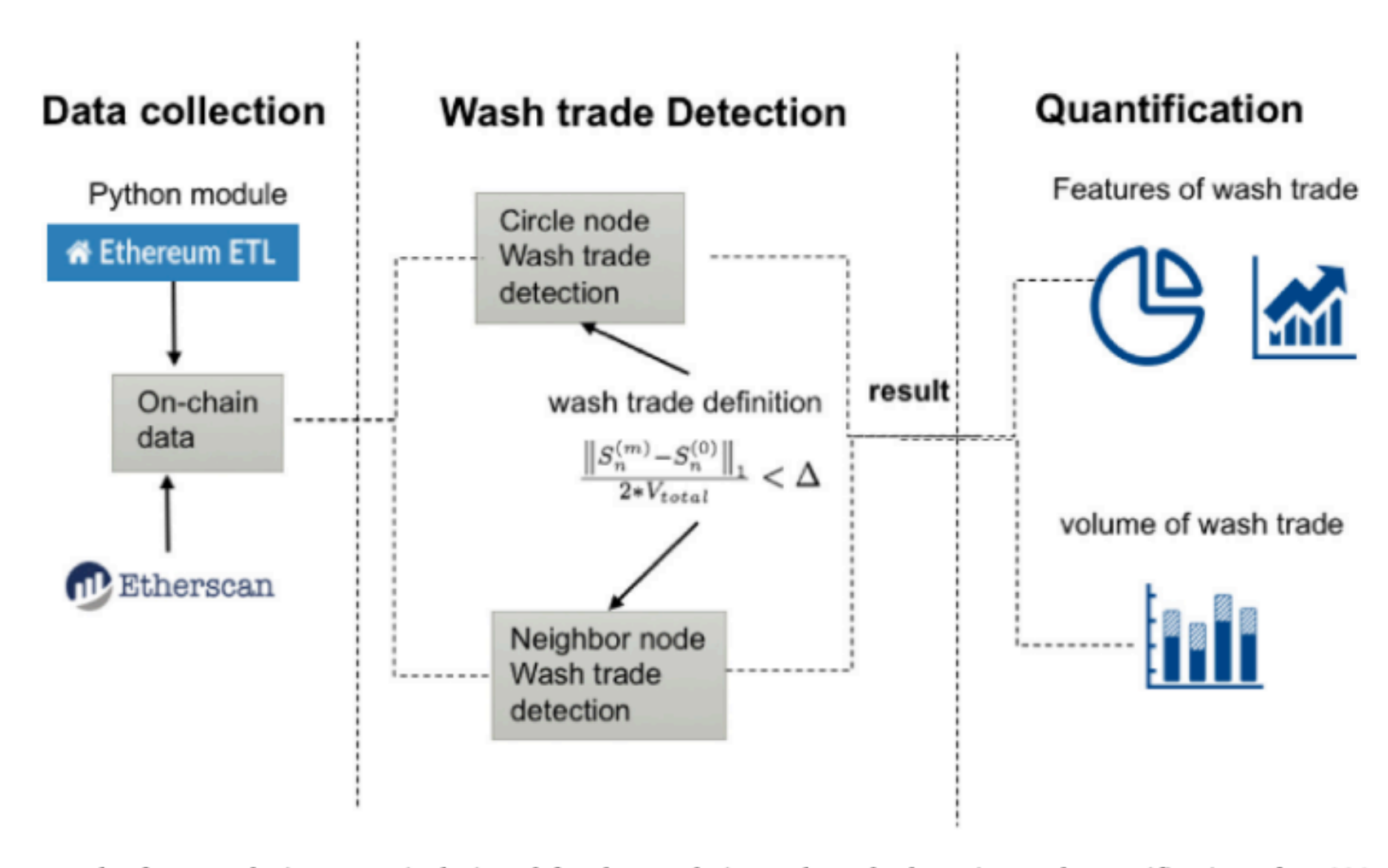

Fig. 1. The framework of our analysis system is designed for the on-chain wash trade detection and quantification of ERC20 cryptocurrency.

首先介绍了区块链技术的起源与发展，指出比特币网络因缺乏图灵完备性难以满足复杂场景需求，而以太坊通过智能合约实现了图灵完备性，推动了去中心化应用（DApps）和初始代币发行（ICOs）的兴起，基于ERC20标准的代币在以太坊平台上大量涌现，截至论文撰写时已达456,761种，2021年10月加密货币市场市值超过2.5万亿美元。随后，文章梳理了加密货币经济学的主要研究方向，包括货币经济学、信息效率、价格波动等，同时强调市场操纵行为如对敲交易因区块链的匿名性和监管缺失成为重要研究课题——对敲交易通过同一实体或合谋团体控制的地址进行反复交易，虚增交易量以营造市场活跃假象，误导投资者做出错误判断，在DeFi项目蓬勃发展、ERC20代币大量发行的背景下，此类行为对代币价格波动和市场结构影响显著。文章进一步指出现有研究多聚焦于传统金融市场或交易所数据，缺乏针对ERC20代币链上交易数据的高效检测方法，且算法时间与空间复杂度较高，难以实时监测。基于此，论文提出两种算法用于检测对敲交易并量化其规模，通过构建交易图模型和数学定义，首次基于链上数据对ERC20代币对敲交易进行系统性分析，开发工具WTEYE实现特征分析与规模量化。实验结果显示，多数ERC20代币的对敲交易率超过15%，UNI代币超30%的交易被判定为对敲交易，揭示了市场活跃度存在虚假成分，强调恢复真实交易数据对市场监管的必要性。

Data preparation

主要围绕ERC20代币链上交易数据的获取与预处理展开。由于ERC20代币可通过智能合约无监管发行，其交易易被合谋团体操纵以虚增交易量，研究从以太坊浏览器Etherscan和Python工具包Ethereum ETL采集2021年1月1日至3月1日期间的链上交易数据，包含时间戳、交易哈希、区块号、收发地址及交易数量等字段，并按单日交易打包存储。考虑到DeFi（去中心化金融）协议催生的ERC20代币（如LINK、MKR、COMP、CRV、UNI等）在市场中的重要性及对敲交易对其网络结构和价格的显著影响，研究聚焦于与DeFi相关的代币。预处理阶段，为减少计算复杂度并提升效率，剔除了对实验结果影响微小的小交易量数据和边缘节点，仅保留时间戳、收发地址及代币交易数量等核心信息。处理后的数据概览显示，LINK是交易最活跃的代币（交易地址超100万，交易数量超10亿），而MKR的交易地址最少（约15万），为后续对敲交易检测算法提供了高质量的输入数据集。

Wash trade detection

Table 1
Overview of some ERC20 token datasets.

Protocol	ERC20 token	Trade	Address	Quantity
Chainlink	LINK	1,065,435	464,108	1,040,502,181
Maker DAO	MKR	154,216	57,864	2,471,364
Compound	COMP	211,374	108,018	15,001,128
Curve	CRV	187,099	30,430	1,791,333,920
Uniswap	UNI	513,812	221,619	1,050,645,311

根据英国金融行为监管局（FCA）和欧洲证券监管委员会（CESR）的定义，对敲交易（wash trade）是指虽遵循合法交易规则、但由合谋团体故意预先安排的交易行为，其目的是伪造交易量并误导其他交易对手方。在本节中，对敲交易的检测流程被划分为三个步骤：第一步，基于数据结构构建交易图模型；第二步，提出两种算法对敲交易进行定量检测；第三步，依据对敲交易定义对符合条件的交易进行标记。

1. 代币交易图模型

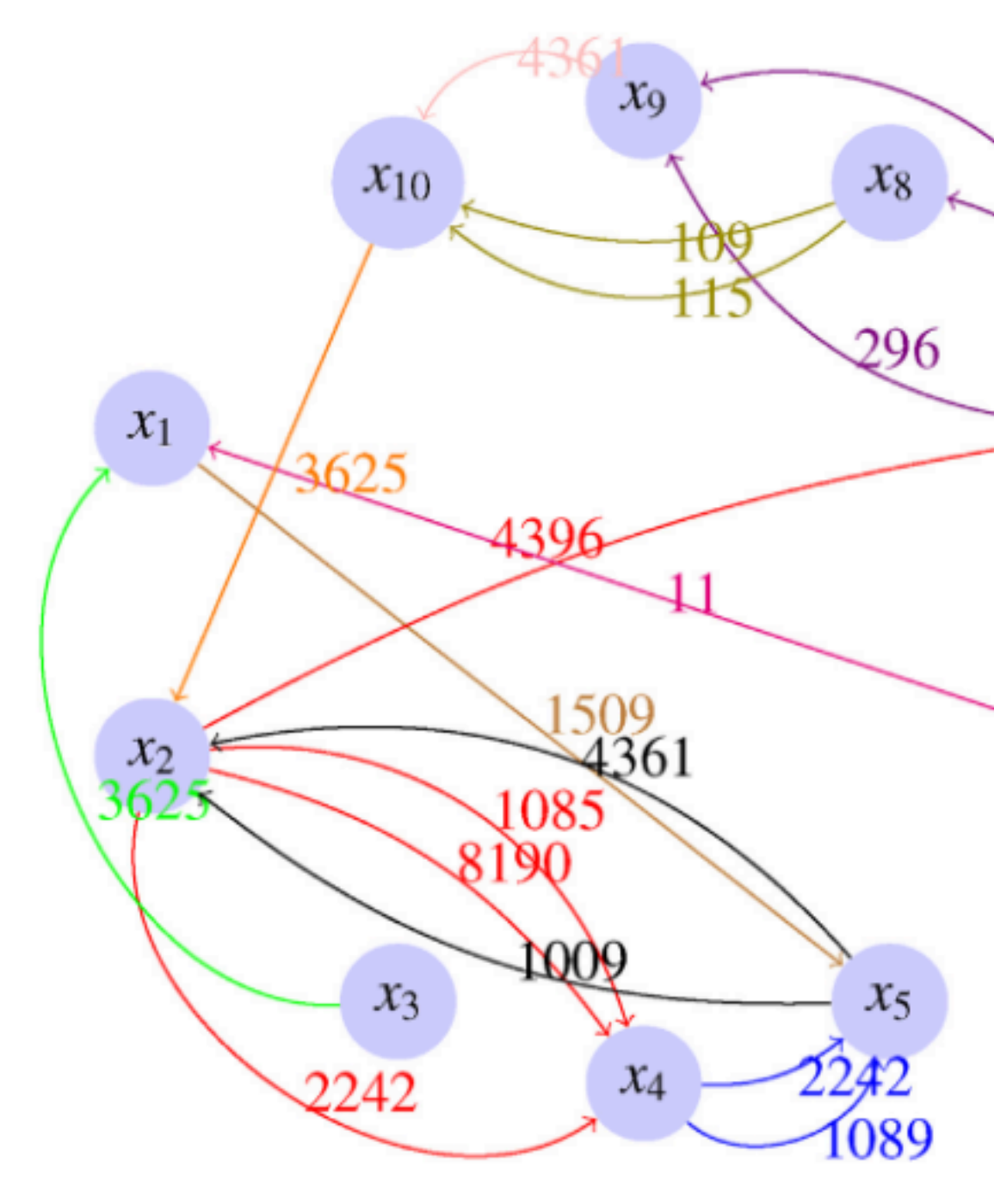

Fig. 2. An example of ERC20 token flow graph.

- 研究定义了一个有向图$G = (V, E)$，其中节点集合V代表交易者地址，边集合 E代表交易记录，每条边的权重为交易的代币数量。通过将交易数据映射为图结构，每个地址对应图中的一个节点，当两地址间发生交易时，创建一条从发送地址指向接收地址的有向边，边权为交易数量。
- 图2以包含10个节点和18条边的代币流动图为例，展示了图模型的具体形式：箭头表示代币流动方向，权重标注交易数量。图中存在多组循环交易如节点(x_2, x_4, x_5)和(x_2, x_7, x_8, x_{10})间的循环，这些循环节点可能属于同一合谋团体，其交易虽营造市场活跃假象，但实际未产生真实的交易需求。
- 该模型为后续对敲交易检测算法提供了基础框架，通过分析图中节点间的连接关系和交易流模式，识别具有循环特征或异常流动的交易组合，为定量检测对敲交易奠定了数据结构基础。

2. 地址状态和对敲交易的定义

首先，研究假设合谋团体控制的地址数量为n，其初始状态集表示为$S_n^{(0)}$ = $\{A_1, A_2, ..., A_n\}$，其中A_i代表地址i的初始代币持有量，每笔交易tx = $\{seller_address, buyer_address, value\}$会改变地址状态，例如地址1向地址2转移100枚代币后，状态集更新为$S_n^{(1)}$ = $\{A_1$ −100, A_2+100, ..., $A_n\}$，反映代币在地址间的流动变化。
针对对敲交易，合谋团体通过由m笔交易组成的交易列表tx_list = $\{tx_1, tx_2, ..., tx_m\}$虚增总交易量$V_{total}$ = $Sum[tx_1.value, tx_2.value, ..., tx_m.value]$，同时通过循环交易使地址状态接近初始状态以减少经济损失。研究定义交易后的终端状态集为$S_n^{(m)}$，若状态集变化与总交易量满足$\frac{||S_n^{(m)}-S_n^{(0)}||1}{2\times V\text{total}} < \Delta$（其中$\Delta$为有效所有权变化率阈值，实验中设为10%），则判定这些交易为对敲交易。该条件表明，尽管交易次数和交易量被放大，但地址间代币的净转移量极低，实质所有权未发生有效变更，符合对敲交易“伪造交易量、误导市场”的核心特征，为后续检测算法提供了量化判定依据。

3. 检测方法

3.1 循环节点对敲交易检测算法（深度优先）

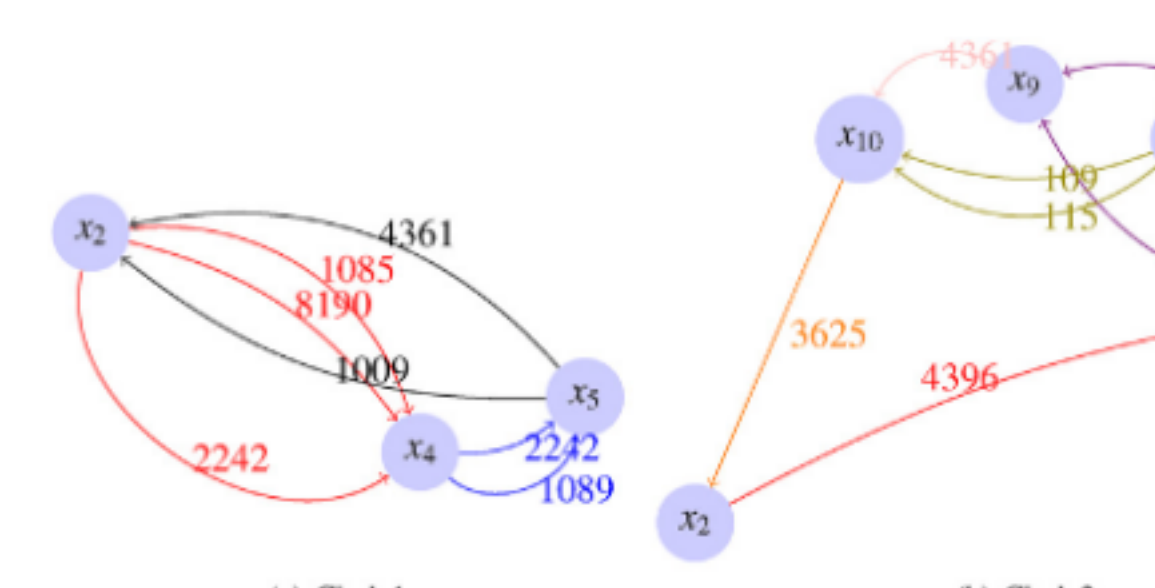

Fig. 3. ERC20 token flow subgraph of circle nodes, which are detected by the algorithm we propose.

- 循环节点对敲交易检测算法旨在通过图遍历识别合谋团体控制的循环交易模式。该算法以代币流动图（包含所有活跃地址及交易列表）为输入，利用深度优先搜索DFS遍历图中节点，寻找形成循环的节点集合（如节点组$\{x_2, x_4, x_5\}$）。若循环节点间的交易满足对敲交易定义（即多次交易后地址状态变化与总交易量的比值低于10%的阈值），则将这些交易标记为可疑对敲交易。
- 算法流程为：构建节点–邻居映射表，从每个关键节点出发进行DFS搜索，记录循环路径；对每个循环节点组，计算其交易后的地址状态变化率R，若$R < \Delta$（实验中Δ=10%），则判定为对敲交易。尽管图2中的示例循环因状态变化率超过阈值未被标记，但该算法为检测具有循环特征的对敲交易提供了系统化方法，后续可通过筛选关键节点优化效率。

3.2 邻居节点对敲交易检测算法（广度优先）

Table 2
Key node map.

Key node	Neighbor node
x_1	$\{x_5\}$
x_2	$\{x_4, x_7\}$
x_3	$\{\}$
x_4	$\{x_5\}$
x_5	$\{x_2\}$
x_6	$\{x_1\}$
x_7	$\{x_8, x_9\}$
x_8	$\{x_{10}\}$
x_9	$\{x_{10}\}$
x_{10}	$\{x_2\}$

- 邻居节点对敲交易检测算法旨在通过迭代扩展节点邻居关系，更全面地搜索链上对敲交易。该算法以代币流动图中的“关键节点”为起点，将当前节点及其所有邻居节点逐步加入检测列表，通过反复查找邻居节点直至无新节点可扩展，形成包含关键节点及其多层邻居的节点集合。若这些节点间的交易满足对敲交易定义（即多次交易后地址状态变化率低于10%的阈值），则将相关交易标记为对敲交易。
- 算法流程为：输入关键节点及其邻居节点映射表，从每个关键节点出发，逐层扩展邻居节点形成动态节点列表，对列表内节点间的交易进行状态变化率计算和条件判断。相较于循环节点检测算法，该方法覆盖范围更广（如图2中除孤立节点x_3和x_6外的所有节点均被遍历），但时间和空间复杂度更高。因此，研究提出过滤边缘节点和小交易量数据以优化效率，确保算法在大规模地址场景下的实用性。该算法通过邻居节点的迭代检测，弥补了循环节点检测的局限性，为识别合谋团体在局部网络中的对敲交易提供了更全面的解决方案。

Experiment results and quantification

1. 对敲交易特征

通过分析代币交易记录的结构，研究得出了有效监控对敲交易的三个特征。操纵者为将利润损失风险降至最低，往往会尽可能快速地匹配对敲交易，因此交易时间间隔和单笔交易数量可视为对敲交易的两个重要特征。此外，合谋客户的数量也是对敲交易的特征之一。关于这些特征的详细信息及实验结果如下文所述。

1.1 单次对敲交易量

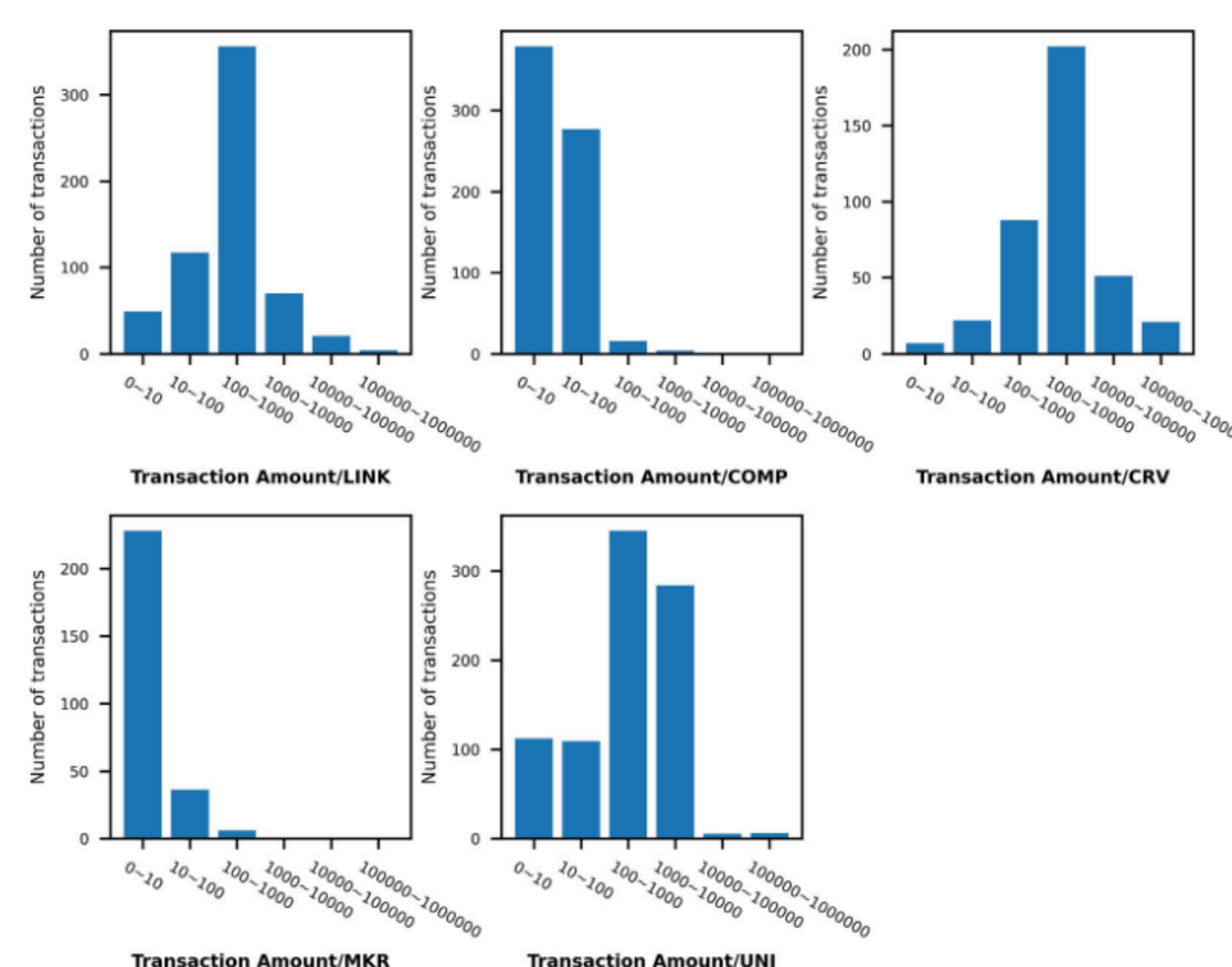

Fig. 4. Feature1: the quantity distributions of a single wash trade, including five kinds of ERC20 tokens.

操纵者为快速虚增交易量，倾向于通过特定规模的交易实现对敲，因此研究聚焦单笔对敲交易数量的分布规律。基于2021年1月1日至1月2日的链上数据，实验发现不同代币的单笔对敲交易数量呈现显著差异：LINK的多数交易数量q_t集中在100至1000之间，CRV和UNI的q_t多为1000至10000，而COMP和MKR的q_t普遍小于10。这一差异与代币价格高度相关——例如LINK价格约11 USDT，MKR约600 USDT，COMP约140 USDT，CRV约0.6 USDT，UNI约4.7 USDT，单笔对敲交易数量与代币价格成反比，符合公式$q_t = \frac{10000USDT}{Price_{token}}$。该特征表明，对敲交易的单笔规模通常围绕“10000美元等值代币数量”波动，为检测系统提供了关键参数参考，可通过设定与代币价格相关的阈值，高效筛选可疑交易，提升检测效率。

1.2 时间间隔

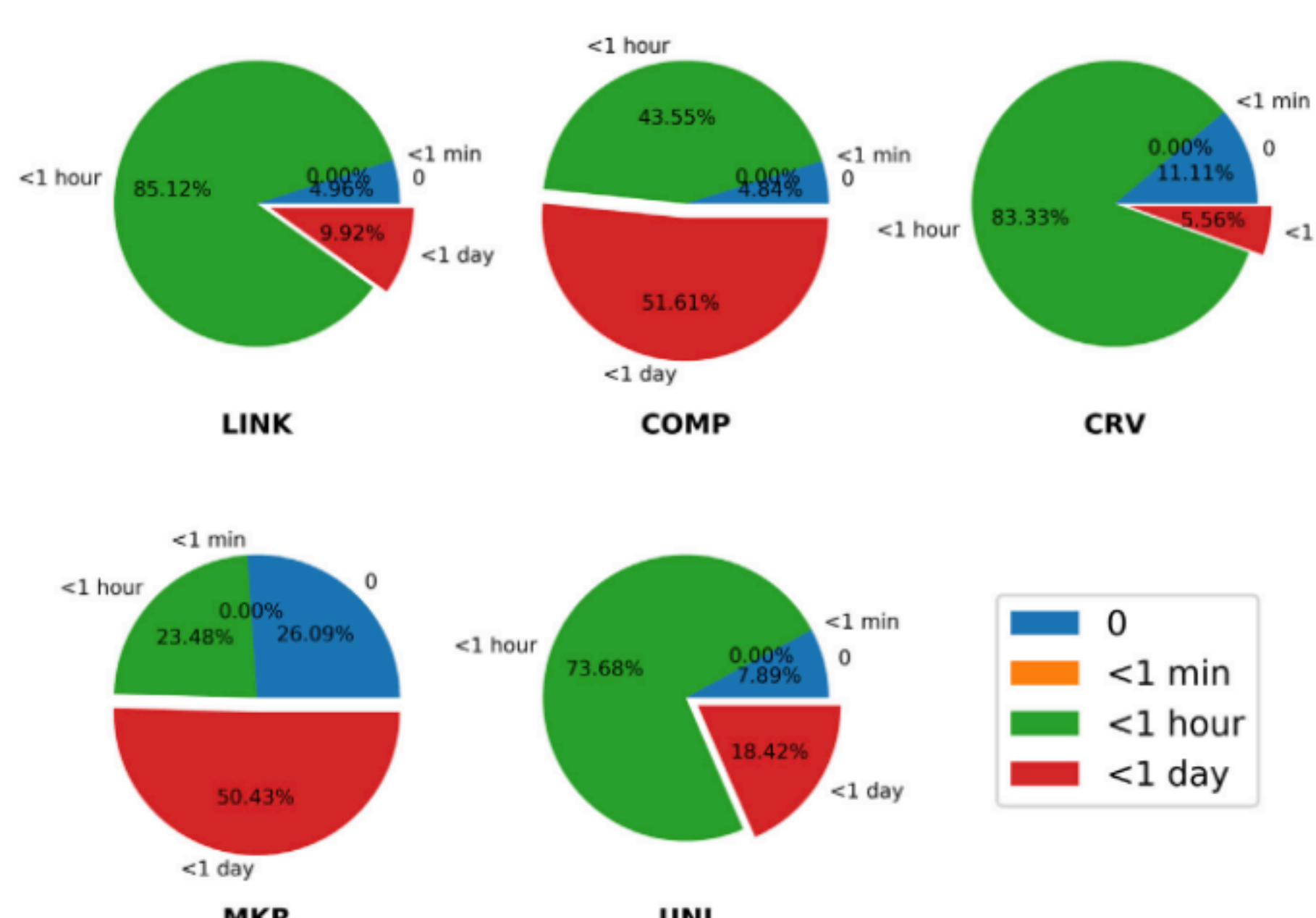

Fig. 5. Feature2: the time interval of wash trade, including five kinds of ERC20 tokens.

由于加密货币价格波动剧烈，对敲交易操纵者为降低价格波动带来的潜在损失，倾向于在短时间内完成交易，甚至将多笔交易打包至同一块中以最小化时间差。实验基于2021年1月1日至3月1日的链上数据，结果显示：除COMP和MKR外，多数代币的对敲交易时间间隔小于1小时，其中LINK代币85.12%的对敲交易间隔在1小时内，MKR代币26.09%的对敲交易间隔为零（即被打包至同一块）。与去中心化交易所相关的代币（如LINK、UNI、CRV）对敲交易更为活跃，时间间隔普遍较小，而与去中心化交易所无关的代币（如COMP、MKR）多数对敲交易在1小时内完成。该特征表明，对敲交易的时间间隔与代币所属的市场场景（如是否关联DeFi）密切相关，通过缩小时间检测窗口（尤其是针对交易活跃的代币），可更精准地捕捉对敲交易的时间模式，提升检测效率。

1.3 合谋客户数量

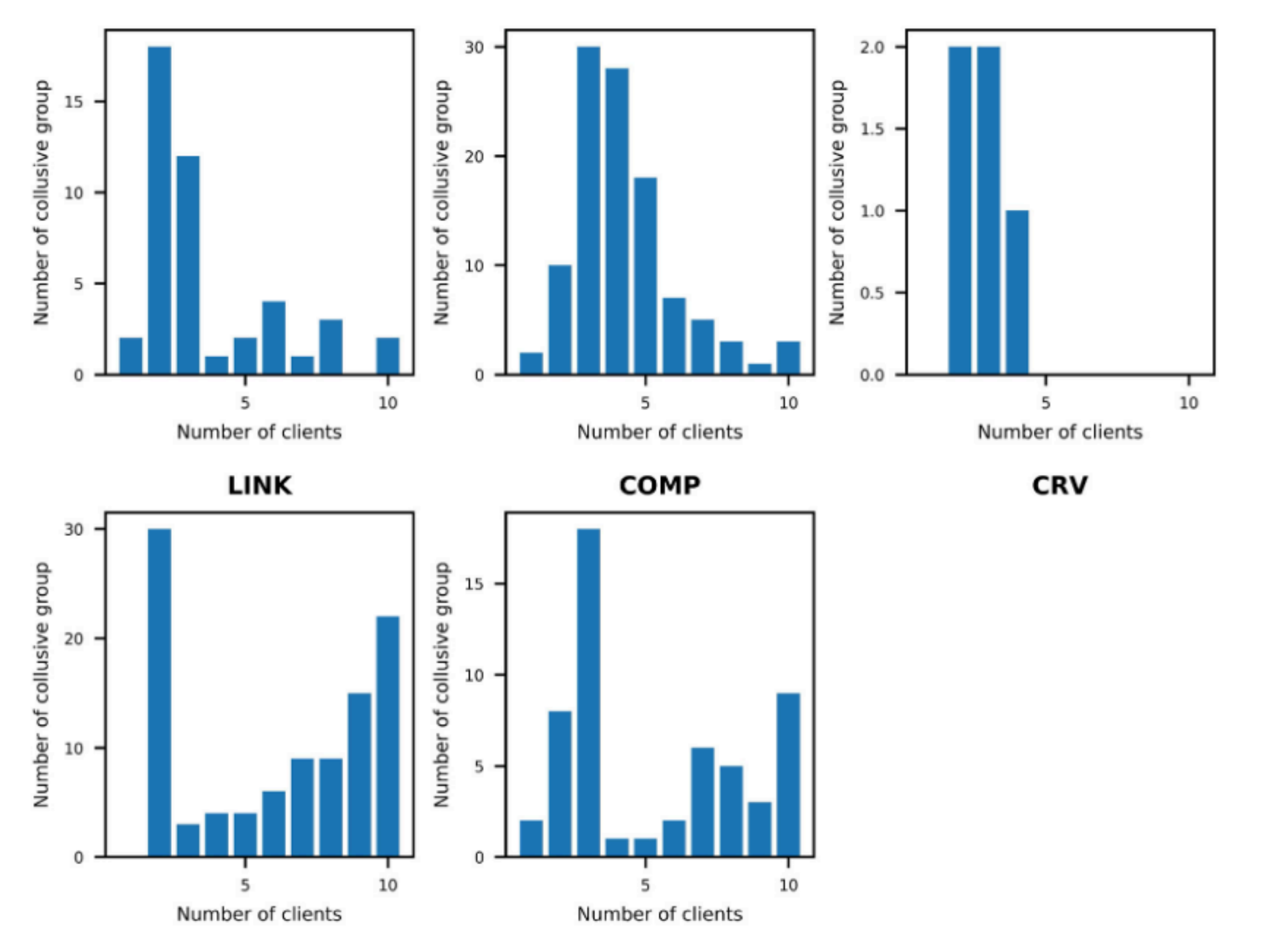

Fig. 6. Feature3: the number of collusive clients of wash trade, including five kinds of ERC20 tokens.

区块链规则限制自我交易，因此合谋团体需通过至少两个地址实施对敲交易，但过多地址可能增加经济损失风险。实验通过对五种ERC20代币的分析发现，除CRV外，多数对敲交易由3至4个合谋客户发起，这些地址通常属于同一实体或合谋团体。图6结果显示，LINK、COMP、MKR、UNI的对敲交易中，合谋客户数量集中在3–4个，而CRV的分布相对分散。该特征表明，对敲交易多发生在小范围地址集群内，因此检测时可将范围缩小至相关地址数量少于20的交易，通过聚焦少数地址间的异常交易模式，有效提升对敲交易的识别效率，减少无关数据干扰。

2. 对敲交易量化

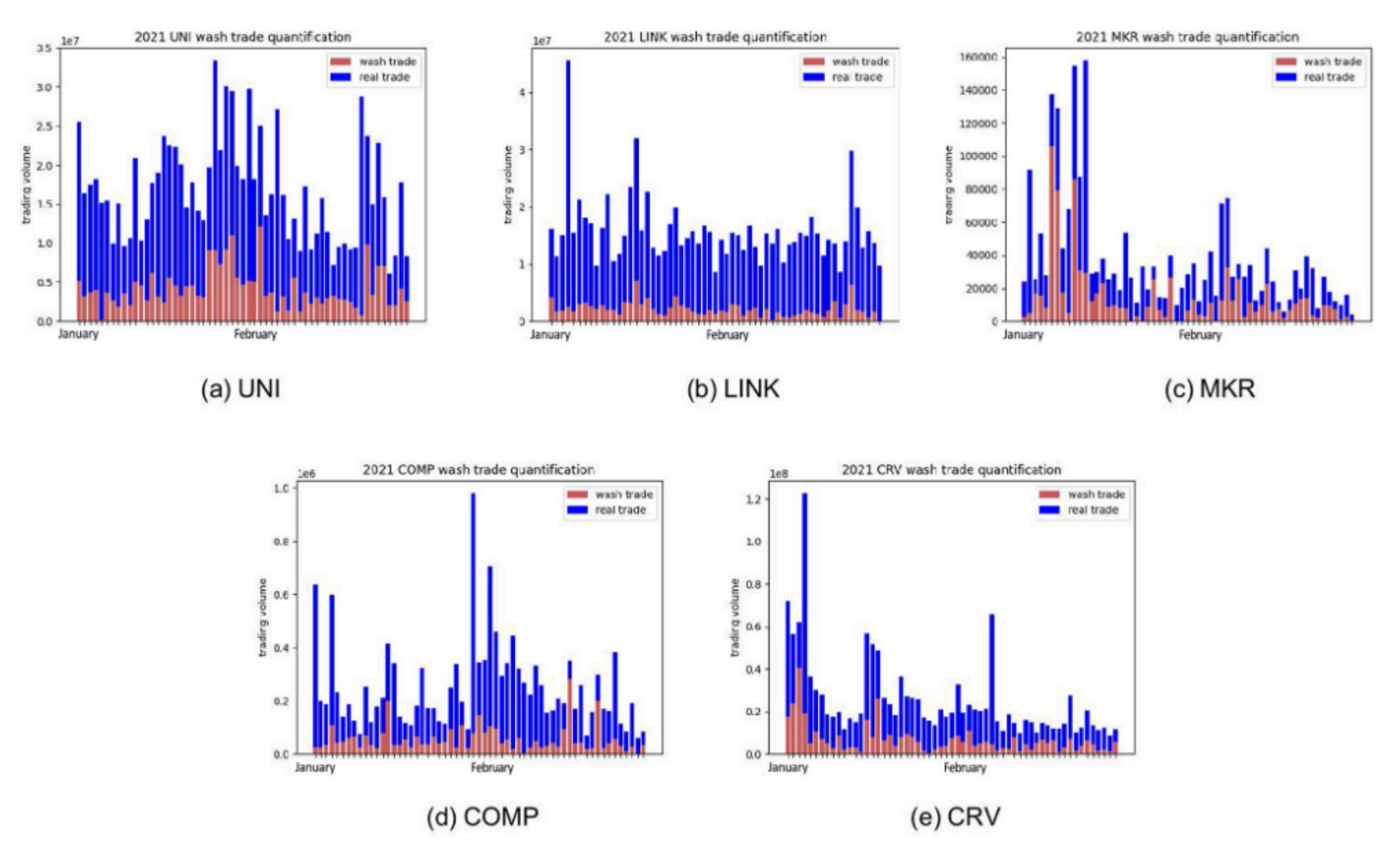

Fig. 7. The quantification of wash trade for five kinds of ERC20 tokens from January 1, 2021 to March 1, 2021.

研究收集了2021年1月1日至3月1日的链上数据，时间窗口设定为1天，通过前期提出的检测算法标记对敲交易并统计其占比。结果显示，对敲交易在各代币中普遍存在，且对敲交易数量与总交易量呈正相关——当对敲交易增加时，总交易量同步上升，表明对敲交易对市场方向具有引导作用，是常见的市场操纵手段。
具体而言，LINK的对敲交易率低于其他代币，而UNI代币超过30%的交易被判定为对敲交易，显示其市场操纵程度较高。实验结果证实，多数ERC20代币的对敲交易率超过15%，揭示出市场活跃度存在显著虚假成分。该量化分析为识别代币交易的真实性提供了数据支持，进一步证明恢复真实交易数据对市场监管的必要性，也为后续开发对敲交易检测工具（如WTEYE）提供了实证依据。

3. 一种用于对敲交易检测的工具

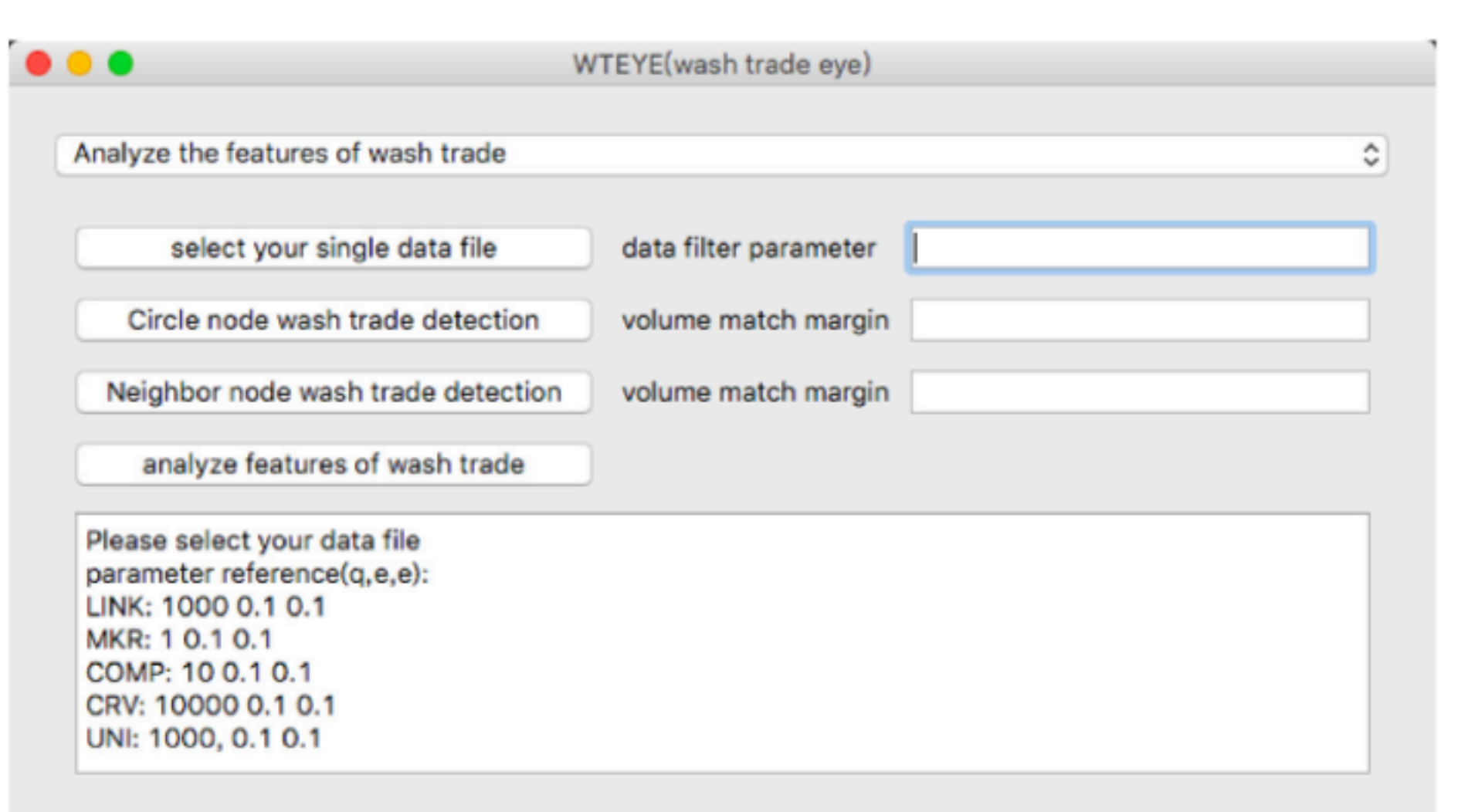

Fig. 8. WTEYE page 1: analyze the features of wash trade.

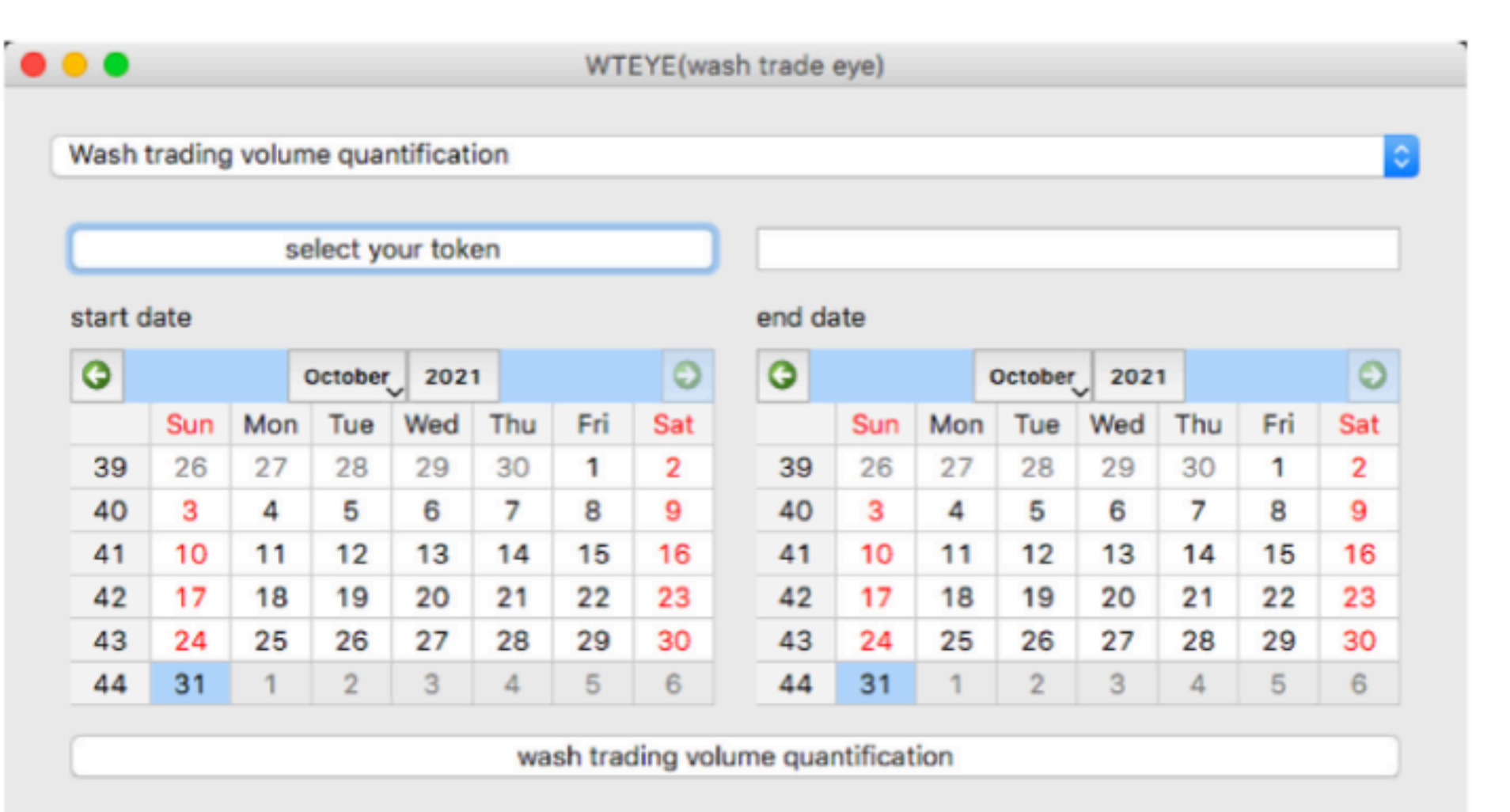

Fig. 9. WTEYE page 2: quantify the percentage of wash trade.

工具包含两个功能页面：**页面1**用于分析对敲交易特征，用户输入指定代币的单日交易数据及参数（如单交易规模阈值、时间间隔阈值），WTEYE通过图8所示界面标记对敲交易并以图表展示其特征（如单笔交易数量分布、时间间隔占比、合谋客户数量），同时提供不同代币的参数参考（如LINK的单交易规模参考值为1000，UNI为1000）；**页面2**用于量化对敲交易规模，用户选择检测时间段和目标代币后，工具显示该期间内的对敲交易量及真实交易量（如图9所示时间选择和结果展示布局），帮助用户直观识别市场活跃度中的虚假成分。
此外，WTEYE具有扩展性，可适配交易数据结构与ERC20代币一致的其他市场（如传统金融市场），通过输入同类交易记录即可实现对敲交易检测，为多场景下的市场监管提供了通用化工具支持。

Conclusions and future work

研究通过构建ERC20代币链上交易的图模型与地址状态数学定义，提出循环节点和邻居节点两种对敲交易检测算法，首次实现了对敲交易的链上量化检测。实验分析显示多数ERC20代币对敲交易率超过15%，其中UNI代币超30%的交易被判定为对敲交易，证实市场活跃度存在显著虚假成分，并提炼出对敲交易单笔规模与代币价格负相关、时间间隔多集中在1小时内、合谋客户数量多为3–4个地址等核心特征，开发的可视化工具WTEYE可支持特征分析与规模量化，为市场监管提供数据支撑。未来研究计划从三方面拓展：针对以太坊2.0等大规模链上数据优化算法效率以降低复杂度，将检测方法应用于比特币、NFT或传统金融市场，结合图神经网络等机器学习技术挖掘深层合谋模式并引入监管沙盒机制实时监测异常交易，以提升检测精准度并拓展应用场景。研究揭示了加密货币市场对敲交易的普遍性与危害性，证明链上数据分析在市场操纵检测中的关键作用，为构建透明化监管体系奠定了方法基础。