Viralização de Questionários Online: desafios e oportunidades

Maurício El Uri[1(*)], Rafael Kreutz[2(*)], Maurício Fiorenza[1(*)], Diego Kreutz[1(*)], Thiago Escarrone[1], Daniel Temp[1], Vinicius Nunez[1], Rodrigo Mansilha[1]

[(*)] **Nota: os 4 (quatro) primeiros autores contribuíram de maneira equivalente no desenvolvimento do trabalho.**

[1]Laboratório de Estudos Avançados (LEA)
Programa de Pós-Graduação em Engenharia de Software (PPGES)
Universidade Federal do Pampa (UNIPAMPA)
mauricioeluri@gmail.com,{NomeSobrenome}@unipampa.edu.br

[2]Grupo de Pesquisa Políticas, Gestão Pública e Desenvolvimento (GPPGPD)
Universidade Estadual do Rio Grande do Sul (UERGS)
rafaelkreutz@gmail.com

Resumo. *Nos últimos anos, temos observado um aumento na quantidade de coletas de dados realizadas através de questionários online. A partir da coleta e análise de 64 questionários online, mostramos como o uso desse recurso é suscetível a usuários maliciosos, que podem agir de forma automatizada no comprometimento da confiabilidade dos dados coletados. Os questionários dos principais sistemas empregados, como Google Forms, foram submetidos a testes básicos de segurança, como personificação e poluição de conteúdo. Os resultados apontam que a maioria dos questionários criados e aplicados online é suscetível a ataques do tipo* sybil*. Além disso, verificamos que a maioria dos sistemas de criação dos questionários online é desprovida de recursos cruciais para algumas pesquisas, como a limitação do grupo de respondentes, a privacidade e a confidencialidade dos dados coletados.*

1. Introdução

Estudos e observações empíricas, como as apresentadas neste trabalho, apontam que a coleta de dados para pesquisas científicas, diagnósticos de mercado, definição de políticas públicas, ações sociais, ou de gestão, tem sido realizada, crescentemente, através de questionários (ou formulários) online [SALVADOR et al., 2020, REGMI et al., 2016, MONTEIRO and SANTOS, 2019, HSU and WANG, 2017]. Um dos aspectos que impulsiona a popularização da utilização de questionários online é o fato de os sistemas, como Google Forms, SurveyMonkey e LimeSurvey, agilizarem e simplificarem o processo de coleta de dados. Entretanto, estudos indicam, de maneira recorrente, diferentes tipos de desafios e limitações na utilização de questionários online, como falta de identificação dos respondentes (ou anonimato), privacidade e confidencialidade no armazenamento de dados sensíveis [VITAK et al., 2016, REGMI et al., 2016, CAMPOS et al., 2011, BUCHANAN and HVIZDAK, 2009, MONTEIRO and SANTOS, 2019].

Na prática, limitações como a falta de identificação dos respondentes podem impactar de forma significativa a confiabilidade dos dados coletados, ou seja, comprometer, inclusive, a validade dos dados das pesquisas realizadas por meio de questionários online [RAAT et al., 2007, BAZAR, 2017, REGMI et al., 2016]. Por exemplo, o uso indiscriminado do Google Forms em pesquisas, que representam 76,6% dos questionários

coletados, aumenta os riscos associados aos resultados (em termos de confiabilidade dos dados) das pesquisas, considerando que usuários maliciosos podem facilmente cometer ataques do tipo *sybil* [DOUCEUR, 2002] para poluir, invalidar ou induzir resultados das pesquisas, como discutimos na Seção 3.

A partir dos desafios e limitações identificados na literatura e de observações empíricas sobre os questionários online, podemos elencar perguntas como: (a) Quais são os tipos mais frequentes de questionários online (*e.g.* pesquisa científica, consulta organizacional, consulta governamental)? (b) Quais são as vulnerabilidades mais comuns, presentes em questionários online atuais? (c) É possível construirmos uma ferramenta automatizada para poluir os dados dos questionários online atuais? (d) Quais são os desafios e as oportunidades com relação à confiabilidade, segurança e privacidade? O principal objetivo neste trabalho é apresentarmos uma resposta inicial a cada uma destas perguntas, através de dados, análises, protótipos e discussões.

As contribuições deste trabalho são: (a) coleta e classificação de questionários online; (b) identificação das principais vulnerabilidades dos questionários coletados; (c) identificação dos desafios mais recorrentemente apontados na literatura e os resultantes das análises técnicas dos questionários coletados; (d) implementação e avaliação de uma ferramenta configurável e automatizada de poluição de dados de questionários online; e (e) análise dos mecanismos de segurança de sistemas de questionários online. As próximas seções apresentam estatísticas (Seção 2) e análise do impacto em potencial de um usuário malicioso (Seção 3) sobre os 64 formulários online coletados, uma comparação dos sistemas de formulários online existentes (Seção 4) e as considerações finais (Seção 5).

2. Estatísticas de questionários online

Entre abril e julho de 2020, período da coleta dos questionários online, os autores receberam 64 formulários por meio de email ou alguma rede social. Os formulários foram manualmente catalogados e analisados segundo quatro critérios: (a) plataforma, (b) abrangência, (c) forma de identificação e (d) finalidade. Como pode ser observado na Figura 1(a), a maioria dos questionários coletados (76,6%) utiliza a ferramenta Google Forms. O restante dos questionários adota o SurveyMonkey (7,8%), entre outros sistemas (15,6%), como iSurvey, LimeSurvey, Qualtrics, REDcap, FormSUS e Cognito Forms. O resultado sugere uma atual predominância da plataforma Google Forms.

Com relação à abrangência (Figura 1(b)), os formulários online foram classificados conforme suposto âmbito de divulgação das coletas de dados: nacional (57,8%), institucional (25%), internacional (9,4%) ou regional (7,8%). Esse critério levanta dúvidas relacionadas ao escopo pretendido do estudo que motivou cada questionário. Por exemplo, pesquisas futuras podem tentar responder a seguinte pergunta: é possível oferecer alguma garantia sobre a abrangência de uma coleta de dados realizada através de um questionário online?

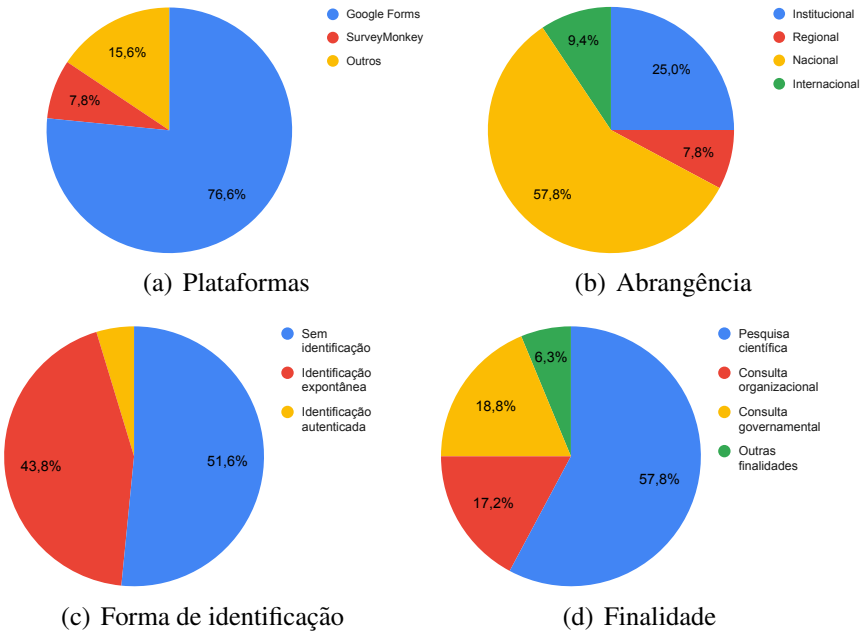
A Figura 1(c) apresenta a classificação dos questionários em quatro possíveis níveis de identificação dos respondentes: sem identificação (51,6%), identificação espontânea (43,8%) ou identificação autenticada (4,7%). No primeiro grupo se enquadram aqueles questionários que não coletam e não solicitam nenhum tipo de dado que permita a identificação dos respondentes (*e.g.*, sob alegação de preservação de anonimato).

No segundo grupo foram enquadrados aqueles questionários que solicitam algum tipo de informação específica do respondente que pode permitir sua identificação, implícita ou explicitamente, como nome, endereço de email, número de CPF, matrícula de aluno e número de telefone. No terceiro grupo se enquadram os questionários que exigem algum processo de autenticação preliminar, como conta institucional combinada com algum fator de identificação, como login e senha.

É importante destacar que 95% dos questionários online coletados não identifica o público alvo da coleta de dados. Nesse tipo de questionário, especula-se que um usuário malicioso poderia ser capaz de realizar um ataque do tipo *sybil* [DOUCEUR, 2002] para submeter uma quantidade não desprezível de respostas e, assim, poluir e, eventualmente, invalidar os resultados da respectiva pesquisa. Além disso, um usuário malicioso pode usar dados coletados na Internet para interferir nos resultados para obter algum tipo de vantagem (veja discussão na Seção 3).

Por fim, os formulários foram classificados de acordo com a finalidade da respectiva pesquisa, como mostra a Figura 1(d), em: pesquisa científica e publicações (57,8%), (b) consulta organizacional (17,2%), (c) consulta governamental (18,8%) ou (d) outras finalidades (6,3%). As consultas são desenhadas para apoiar tomada de decisão (particularmente relacionada à pandemia no grupo estudado) em níveis organizacionais (como instituições de ensino) ou governamentais (*e.g.*, cidade, região, estado ou país). São exemplos de questionários com outras finalidades aqueles voltados ao levantamento de demandas de mercado, parecer sobre atividades ou projetos e compra de produtos.

Figura 1. Estatísticas dos formulários coletados e analisados



(a) Plataformas (b) Abrangência

(c) Forma de identificação (d) Finalidade

Os resultados apresentados permitem tecer uma discussão preliminar sobre desafios e oportunidades de pesquisas relacionados à segurança e privacidade de questionários online. Os números sugerem que a pandemia acelerou o processo de "digitalização" da metodologia conhecida como *survey*, à exemplo do que ocorreu em outras atividades humanas, como tele medicina, tele trabalho e tele ensino. Nesse processo de digitalização de

questionários, observamos empiricamente que os usuários diretos (elaboradores) e indiretos (respondentes) têm preferido o sistema Google Forms, que pode servir como padrão *de facto* a ser investigado. Os resultados obtidos através dos sistemas de questionário online atuais podem estar sendo impactados por usuários leigos, que podem configurar ou utilizar de maneira inadequada os sistemas dada sua aplicabilidade, e por usuários maliciosos, que pretendem obter vantagem ou não. A quantidade expressiva de estudos científicos baseados em dados coletados através de questionários online sustentam a relevância e a emergência do assunto para a comunidade científica.

3. O impacto (em potencial) de um usuário malicioso

Usuários maliciosos desenvolvem e utilizam frequentemente ferramentas automatizadas para realizar ataques. Como mais de 95% dos questionários online coletados não utilizam qualquer tipo de identificação (*e.g.*, token único por respondente) ou autenticação para o controle de acesso da população alvo, pergunta-se: Qual é o impacto de um ataque automatizado sobre os dados de um questionário online?

Para responder a essa pergunta, um atacante pode coletar dados de pessoas e instituições disponíveis publicamente na Internet[1], em sites de prefeituras e outras instituições públicas e privadas. Com o objetivo de simular o comportamento de um atacante, para fins exclusivos de pesquisa acadêmica, verificamos que, de fato, é possível coletar dados publicados (e abertos) em sites oficiais de instituições públicas e privadas e, consequentemente, organizar em uma base de dados quantidades expressivas (*e.g.* vários milhares de registros) de informações sobre pessoas e instituições, incluindo nome completo, CPF, RG, email, data de nascimento, endereço, telefone, profissões, cargos e matrícula de estudantes de instituições de ensino.

Também implementamos um programa de preenchimento automatizado de dados de questionários online, denominado de PAFO (Poluidor Automatizado de Formulários Online). O PAFO recebe as seguintes entradas: (a) um arquivo CSV contendo as entradas (ou *inputs*) do questionário online; (b) um arquivo CSV contendo as respostas para as questões textuais e objetivas dos formulários; (c) o endereço Web (ou URL) do questionário online. O arquivo CSV do item (b) é criado por um segundo programa, que utiliza uma base de dados externos para gerar respostas de maneira pseudo-aleatória. Ao aplicarmos o PAFO, alcançamos personificação de respondentes e poluição de conteúdo em formulários coletados e que não possuem autenticação dos respondentes, ou seja, com uma ferramenta deste tipo, um atacante consegue (potencialmente) comprometer a confiabilidade dos dados de 95% dos questionários coletados.

Atualmente, o PAFO é capaz de preencher sequencialmente os formulários usando duas estratégias distintas quanto ao intervalo de tempo entre duas submissões consecutivas: constante e pseudo-aleatória. Na primeira estratégia, o PAFO submete respostas adotando uma frequência aproximadamente constante, maximizando a quantidade de respostas falsas submetidas em um determinado período de tempo. Contudo, a taxa constante pode ser utilizada para detectar e filtrar as respostas falsas durante a análise dos dados. Para contornar essa limitação, na segunda estratégia, o PAFO adota tempos pseudo-aleatórios (configuráveis) entre respostas. Isso tende a gerar uma poluição de respostas

[1]Como por exemplo, resultados de concursos públicos, licitações e contratações publicados por prefeituras, secretarias de estado e outros órgãos de governo.

mais próxima do comportamento humano e mais difícil de ser detectada e removida da coleta de dados em comparação com o ataque da primeira estratégia.

Nas duas estratégias, a taxa média pode ser ajustada para aumentar as chances de que as respostas do PAFO sejam intercaladas com as respostas de usuários não maliciosos. Isso deve levantar menos suspeitas e tornar mais difícil a distinção entre repostas honestas e repostas falsas.

4. Sistemas de questionários online

Os seguintes dez sistemas de criação e coleta de dados de questionários online foram selecionados a partir dos questionários coletados pelos autores e de sites especializados em avaliação e recomendação de ferramentas desse tipo: Google Forms, SurveyMonkey, Qualtrics, LimeSurvey (web), LimeSurvey (instalado localmente), Typeform, Microsoft Forms, Paperform, FormSite e Hubspot Free Online Form Builder. A Tabela 1 resume as quatro principais características avaliadas, incluindo controle de acesso (identificação única e método de autenticação) e funcionalidades (anonimização de dados e identificação de grupos).

Tabela 1. Análise dos sistemas de criação de questionários online

	Controle de acesso		**Funcionalidades**	
Sistema	**Identificação única**	**Método de autenticação**	**Permite Anonimização de Dados**	**Permite identificação de grupos**
Google Forms	Não	Por usuário	Não	Não
SurveyMonkey,Qualtrics	Sim	Por grupo	Sim	Sim
LimeSurvey-web	Sim	Sem autenticação	Sim	Sim
LimeSurvey-instalado	Sim	Por usuário	Sim	Sim
Microsoft Forms	Não	Por usuário	Não	Não
Paperform, FormSite, Typeform, Hubspot	Não	Sem autenticação	Não	Não

Vale ressaltar que apenas o Google Forms, o LimeSurvey (instalado localmente), e o Microsoft Forms são gratuitos e sem limite. Talvez, a gratuidade seja um dos fatores que explica a popularidade do Google Forms, que, além disso, apresenta interface familiar e integrada com o conjunto de ferramentas Google Docs.

Com relação ao controle de acesso, todos os sistemas avaliados possibilitam acesso anônimo ao preenchimento dos questionários. Apenas três sistemas (SurveyMonkey, Qualtrics, LimeSurvey) possuem algum tipo de recurso para acesso através de identificação única, como por exemplo, uma URL contendo um token único para cada um dos respondentes. Esta opção pode ser importante quando o questionário é direcionado para um grupo limitado ou conhecido de respondentes. Como pode ser observado na tabela, apenas esses três sistemas oferecem a funcionalidade de criação e identificação de grupos de respondentes.

No caso do Google Forms, LimeSurvey (instalado localmente) e Microsoft Forms, há ainda a opção de forçar a autenticação do respondente. Por exemplo, no caso do Google Forms institucional, via GSuite, é possível limitar o acesso ao questionário online aos usuários autenticados da instituição. Entretanto, não é possível delimitar grupos específicos, como docentes, técnicos, estudantes, ou outros grupos quaisquer. Cada usuário da instituição poderá responder apenas uma única vez o formulário se a opção de autenticação e resposta individualizada for explicitamente habilitada pelo criador do

formulário. Se o criador do questionário online apenas criar e aplicar no contexto institucional, uma única credencial vazada é o suficiente para um atacante poluir (*e.g.*, utilizar uma ferramenta com o PAFO) os dados da pesquisa.

Com relação à funcionalidade de anonimização de dados, apenas o SurveyMonkey, o Qualtrics e o LimeSurvey permitem ao mesmo tempo identificação individualizada dos respondentes e anonimização das respostas, dos dados coletados. Por exemplo, a URL do questionário contém um token único para cada respondente, que é enviado por email. Entretanto, o token único não permite identificar o respondente, pois ele é algo como uma cadeia de caracteres pseudo-aleatória, criada de forma única para cada combinação de questionário/respondente. O criador do formulário desconhece a associação entre o token e o respondente. Este é um recurso interessante para aumentar a confiabilidade dos dados, pois o questionário será aplicado a um grupo conhecido e limitado de respondentes. Entretanto, um administrador (malicioso) do sistema tem acesso à informação que relaciona o respondente ao respectivo token em tempo de execução do sistema.

Outra funcionalidade importante de um sistema de questionários online é registrar o tempo de resposta por página e por questionário. Através do cálculo do alfa de Cronbach, utilizando os tempos de resposta, é possível identificar e classificar qualitativamente os respondentes [MONTAG and REUTER, 2008]. Entretanto, ferramentas de ataque como o PAFO podem ser alimentadas por estatísticas de questionários online, como o alfa de Cronbach, potencializando classificar qualitativamente as respostas da ferramenta como respondentes reais. Dos sistemas avaliados, apenas o SurveyMonkey e o LimeSurvey coletam o tempo de reposta por página. Outras soluções, como Typeform, Qualtrics e FormSite coletam apenas o tempo total de resposta do questionário.

5. Considerações finais

Neste trabalho, apresentamos uma análise de 76,6% questionários online e demonstramos como um atacante pode poluir, de maneira automatizada, o conteúdo de mais de 95% deles, utilizando uma ferramenta como a PAFO, proposta e desenvolvida pelos autores. Além disso, selecionamos também dez sistemas populares de criação de questionários online e analisamos as principais características, como controle de acesso e funcionalidades de anonimização de dados e identificação de grupos.

Entre os trabalhos e desafios futuros, podemos citar: (a) ampliar a análise em termos da quantidade de questionários online coletados e analisados; (b) investigar outras plataformas de questionários online; (c) realizar um estudo mais amplo dos desafios técnicos e não-técnicos da coleta de dados através de questionários online; (d) dar continuidade ao desenvolvimento do PAFO, adicionando estatísticas como a de alfa de Cronbach; e (e) investigar, no contexto de questionários online, o impacto e aplicabilidade das novas tecnologias que objetivam assegurar a privacidade e a segurança de dados dos usuários (*e.g.* novas arquiteturas de sistemas propostas ou avaliadas recentemente [ANTUNES et al., 2018, ISAAK and Hanna, 2018, FRITSCH et al., 2019, CHENTHARA et al., 2019]) e algoritmos de cifra avançados para garantir a confidencialidade dos dados (*e.g.*, apenas o criador da pesquisa e os respondentes podem ter acesso aos dados, *i.e.*, um administrador do sistema não deve ter acesso aos dados).

Referências

ANTUNES, F., GARCIA, F., and KREUTZ, D. (2018). SeguraAí: confidencialidade de dados sensíveis com SGX. In *3o Workshop Regional de Segurança da Informação e de Sistemas Computacionais*.

BAZAR, H. A. (2017). Forms distribution algorithm for online examination systems. In *8th International Conference on Information Technology (ICIT)*, pages 123–126.

BUCHANAN, E. A. and HVIZDAK, E. E. (2009). Online survey tools: Ethical and methodological concerns of human research ethics committees. *Journal of Empirical Research on Human Research Ethics*, 4(2):37–48.

CAMPOS, J. A. D. B., Zucoloto, M. L., Bonafé, F. S. S., Jordani, P. C., and Maroco, J. (2011). Reliability and validity of self-reported burnout in college students: A cross randomized comparison of paper-and-pencil vs. online administration. *Computers in Human Behavior*, 27(5):1875 – 1883. Fifth International Conference on Intelligent Computing.

CHENTHARA, S., Ahmed, K., Wang, H., and Whittaker, F. (2019). Security and privacy-preserving challenges of e-health solutions in cloud computing. *IEEE access*, 7:74361–74382.

DOUCEUR, J. R. (2002). The sybil attack. In Druschel, P., Kaashoek, F., and Rowstron, A., editors, *Peer-to-Peer Systems*, pages 251–260, Berlin, Heidelberg. Springer Berlin Heidelberg.

FRITSCH, L., PAZ, G., MACHADO, R., and KREUTZ, D. (2019). Análise do Impacto de Funções Criptográficas no Desempenho do Google BigQuery. In *4o Workshop Regional de Segurança da Informação e de Sistemas Computacionais*.

HSU, H.-Y. and WANG, S.-K. (2017). Using google forms to collect and analyze data. *Science Scope*, 40(8):64.

ISAAK, J. and Hanna, M. J. (2018). User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59.

MONTAG, C. and REUTER, M. (2008). Does speed in completing an online questionnaire have an influence on its reliability? *CyberPsychology & Behavior*, 11(6):719–721.

MONTEIRO, R. L. d. S. G. and SANTOS, D. S. (2019). A utilização da ferramenta google forms como instrumento de avaliação do ensino na escola superior de guerra. *Revista Carioca de Ciência, Tecnologia e Educação*, 4(2):27–38.

RAAT, H., Mangunkusumo, R. T., Landgraf, J. M., Kloek, G., and Brug, J. (2007). Feasibility, reliability, and validity of adolescent health status measurement by the child health questionnaire child form (chq-cf): internet administration compared with the standard paper version. *Quality of Life Research*, 16(4):675–685.

REGMI, P. R., Waithaka, E., Paudyal, A., Simkhada, P., and van Teijlingen, E. (2016). Guide to the design and application of online questionnaire surveys. *Nepal Journal of Epidemiology*, 6(4):640.

SALVADOR, P. T. C. O., Alves, K. Y. A., Rodrigues, C. C. F. M., and e Oliveira, L. V. (2020). Estratégias de coleta de dados online nas pesquisas qualitativas da área da saúde: scoping review. *Revista Gaúcha de Enfermagem*, 41.

VITAK, J., Shilton, K., and Ashktorab, Z. (2016). Beyond the belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages –.