Behavioral consistency in the digital age

Abstract

Efforts to infer personality from digital footprints have focused on behavioral stability at the trait level without considering situational dependency. We repeat Shoda, Mischel, and Wright’s (1994) classic study of intraindividual consistency with secondary data (5 datasets) containing 28,692 days of smartphone usage from 780 people. Using per app measures of ‘pickup’ frequency and usage duration, we found that profiles of daily smartphone usage were significantly more consistent when taken from the same user than from different users ($d > 1.46$). Random forest models trained on 6 days of behavior identified each of the 780 users in test data with 35.8% / 38.5% (pickup / duration) accuracy. This increased to 73.5% / 75.3% when success was taken as the user appearing in the top 10 predictions (i.e., top 1%). Thus, situation-dependent stability in behavior is present in our digital lives and its uniqueness provides both opportunities and risks to privacy.

KEYWORDS: BEHAVIORAL CONSISTENCY, PERSONALITY, DIGITAL FOOTPRINT, INTRAINDIVIDUAL

Statement of Relevance

Whenever people use technology, they leave behind a digital trace that documents their behavior. We used these data to study—at scale—the question of whether people behave consistently in their digital lives, but in a way that is context dependent. By analyzing 28,692 days of smartphone app usage across 780 individuals, we find that it is possible to profile a person's day-to-day use of different apps and show that this profile remains consistent over time. We show that a single day of data from an anonymous user can be matched to the correct originating user's profile with >70% accuracy when success is taken as the user appearing in the first 10 (top-1%) of all candidates. Thus, people show distinctive patterns of digital behavior even when compared to hundreds of other individuals. This has implications for security and privacy in the digital age.

Introduction

In searching for the locus of personality, psychologists theorize that people behave consistently in situations perceived as psychologically equivalent (Mischel, 2004). This ‘interactionist’ account may be expressed using *if...then* statements. If in Situation X, then a person does Behavior A, but if in Situation Y, then that person does Behavior B (Shoda, Mischel, & Wright, 1994). Originally studied in face-to-face interactions through field observations (Shoda, Mischel, & Wright, 1993) and experimental tests (Furr & Funder, 2004), more recent evidence of behavioral consistency has come from studies of our digital lives. Harari et al. (2019) found high consistency in individuals’ use of call, text messages and social media across days, with the use of social applications (‘apps’) being the most consistent. Aledavood et al. (2015) analyzed the call patterns of 24 individuals over an 18-month period and found that the frequency of calls at each hour of the day was distinct and persistent within an individual.

To date, studies of the consistency of digital behavior have focused on stability in general usage (e.g., calls vs. text messages) or stability within a specific app (e.g., phone calls). There has been no consideration of patterns of behavior across apps. Yet, apps extend our social environment in different ways depending on their features and extrinsic factors (Shaw, Ellis, & Ziegler, 2018). User’s self-identities have amalgamated with the technology they use, as self-expression can be enacted digitally from avatars to social media (Belk, 2013). Therefore, each app represents a ‘nominal situation’ to its user because it comprises a unique interface (i.e., setting) and distinguishing features (i.e., activities) (Davidson & Joinson, 2021; Mischel & Shoda, 2010). It can also elicit mood states (Alvarez-Lozano et al., 2014) and often presents psychological features that are characteristic of ‘active ingredients’ (e.g., peer adoration on Twitter, paper rejection on Email). Qualitative analysis shows these active ingredients differ not only when apps serve distinct functions (e.g., productivity vs social), but also when they offer similar functionality such as communication (Nouwens, Griggio, & Mackay, 2017). Quantitative analysis confirms that daily interactions with apps are unique, even for apps that share identical, high-level categorizations including ‘social media’ (see Supplementary Materials). Collectively, this suggests that smartphone apps are psychologically distinct situations.

Thus, if intraindividual stability exists within digital behavior, we should find that users show different levels of engagement (a behavior) with each situation (the apps), but that this pattern of situation level engagement will remain stable across time (i.e., their personality signature). We test this notion with a pre-registered hypothesis that daily profiles of usage behavior across smartphone apps will show higher intraindividual consistency than inter-individual consistency. We use naturally occurring large-*n* data as a complement to prior observational and experimental studies.

Methods

User Data

We combined five pre-existing datasets that recorded smartphone foreground app use for 1,119 users. These usage data comprised time-stamped actions that showed what app was on a user’s screen, on what day, and for how long (see Supplementary Materials for how we collected these data). Age self-reports were available for 913 users and revealed a wide range of ages in years from 18-24 (*n* = 131), 25-34 (*n* = 214), 35-44 (*n* = 245), 45-54 (*n* = 190), and 55+ (*n* = 133). Of the 909 users who reported their gender, 426 were women and 483 were men.

We standardized, cleaned and combined the datasets (see Supplementary Materials for all procedures). To ensure we could examine behavior across multiple days, we excluded users whose record contained less than nine days of data. We then removed the first and last day of data for each user, since data from these days reflects only a partial day of use depending on when the logging app was installed and uninstalled. This left 824 users in the sample, each of whom had 7 or more days of usage data.

Some apps were used by very few users. Including these apps in our analyses could artificially increase our consistency measure because ‘no use’ would appear across many days and many users. Accordingly, we only included apps used by over 25% of the sample (>206 people) that were not system apps (e.g., ‘Android System’). This left the use of 21 apps in our analysis: Calculator, Calendar, Camera, Clock, Contacts, Facebook, Gallery, Gmail, Google Play Store, Google Search, Instagram, Internet, Maps, Messaging, Messenger, Phone (native phone call app), Photos, Settings, Twitter, WhatsApp, and YouTube.

Assessing Consistency

As outlined by Ellis, Davidson, Shaw, and Geyer (2019), smartphone behaviors may be examined at different levels of specificity. One fundamental behavior is positive engagement, the extent a person acts rather than avoids the situation presented by the app. While such a measure ignores more qualitative aspects in how a person engages (e.g., liking or commenting), the variation in engagement behaviors are themselves a consequence of cognitions and affects about the ‘stimuli’ presented by the app (Shaw, Ellis, & Ziegler, 2018). For example, one person may read and respond enthusastically every time they receive a message in a messenger app, while another may ignore the message and glance only briefly at the end of the day. One metric of engagement is the number of daily app pickups (henceforth ‘Pickups’), which measures the number of times a participant engages. A second useful metric is the daily time spent on the app (henceforth ‘Duration’), which is the equivalent of measuring the magnitude of the engagement. By assessing these variables, it is possible to examine *if…then* patterns of behavior of the form, given a Situation X (app), this person will show Y amount of engagement. While some research suggests that daily Pickups will be more consistent than daily Duration behavior (Wilcockson, Ellis, & Shaw, 2018), we tested both frequency and magnitude to reflect different aspects of our behavioral tendency.

We calculated Pickups and Duration for each app across all the days of data available for each user. We removed days of data where none of the 21 apps were used, which may reflect a technical issue with the logging. This process left 44 users without seven full days of smartphone data, so we removed them, leaving 780 users with full Pickup and Duration data. On average, users had 36.80 days of data (total = 28,692 days), with a minimum of 7 and a maximum of 377 (skewness = 4.61). Pickups were the number of times a user accessed each of the 21 apps per day; Durations were how long in seconds each user spent on each of the 21 apps per day.

Our assessment of consistency followed Shoda, Mischel, and Wright's (1994) approach of comparing profiles of behavior across the 21 apps. We first calculated, for each app, the daily mean and standard deviation of Pickups and Duration (separately); this represented a 'normative profile' of the sample's behavior. We then calculated how each of the 28,692 daily cases deviated from this norm by computing standardized scores (specifically, z scores). For each day's data, for each app's score, we subtracted the sample mean and divided it by the sample standard deviation. The resulting 21 standardized values made up a user's behavioral profile of app use for that day. If a particular app had a score above zero in the behavioral profile, this meant that app was used for longer/more times than the sample norm on that day. Since every user had at least seven days of usage data, we created multiple profiles for each user, allowing us to examine the consistency of profiles over time.

Finally, to ascertain whether apps should be analyzed individually or grouped together into types of apps with similar purposes (e.g., social media apps), we analyzed the structure of the daily behavior profiles using Exploratory Factor Analysis (see Supplementary

Materials). When using an 8-factor solution, findings showed that the variance explained by the factors were low (Pickups = .32, Durations = .19) and indicated no clear way to group the apps together. We thus treated the apps as psychologically distinct situations, with unique daily engagement levels, and analyzed them separately (see Supplementary Materials for full procedures).

This research received ethical approval from the Faculty of Science and Technology Research Ethics Committee (FST19002) and the Security Research Ethics Committee. Our analysis plan was pre-registered (see view only link), and the methods and processed data (distributions of coefficients) are available at: see view only link.

Results

Consistent with the approach of Shoda, Mischel, and Wright (1994), we assessed the similarity of users' daily profiles using ipsative correlations (i.e., we calculated Pearson correlations on rank-ordered profile scores). We did this for two daily profiles randomly selected from the same user (within-user pairs) and two daily profiles randomly selected from different users (between-user pairs). There were 411,601,086 unique comparisons in the data (i.e., $n(n\text{-}1))/2$). We calculated ipsative correlations for 10 million randomly selected within-user pairs and 10 million randomly selected between-user pairs (10 million was our computational limit). We repeated these calculations a further 44 times to obtain bootstrapped Confidence Intervals and effect sizes. See our data visualization website for examples of daily profiles alongside a demonstration of how between and within subject profiles were compared to create distributions: https://behaviouralanalytics.shinyapps.io/AppUseProfiles/

Figure 1 presents the distribution of observed correlations for within- and between-user groups for Pickups (right panel) and Duration (left panel). Confirmatory *t*-tests supported our prediction that Pickups would be higher in within-user pairs (*M* = 0.73, *95%CI*

[0.73, 0.73], *SD* = 0.19) compared to between-user pairs (*M* = 0.30, *95%CI* [0.30, 0.30], *SD* = 0.30], Welsh's *t*(17,004,202) = 3797.93, *p* < .001, *d* = 1.70, 95%CI [1.70, 1.70], and that Duration would be higher in within-user pairs (*M* = 0.81, *SD* = 0.16, 95%CI [0.81, 0.81]) compared to between-user pairs (*M* = 0.49, *SD* = 0.27, *95*%CI [0.49, 0.49]), Welsh's *t*(16,010,722) = 3274.61, *p* < .001, *d* = 1.46, 95%CI [1.46, 1.47].

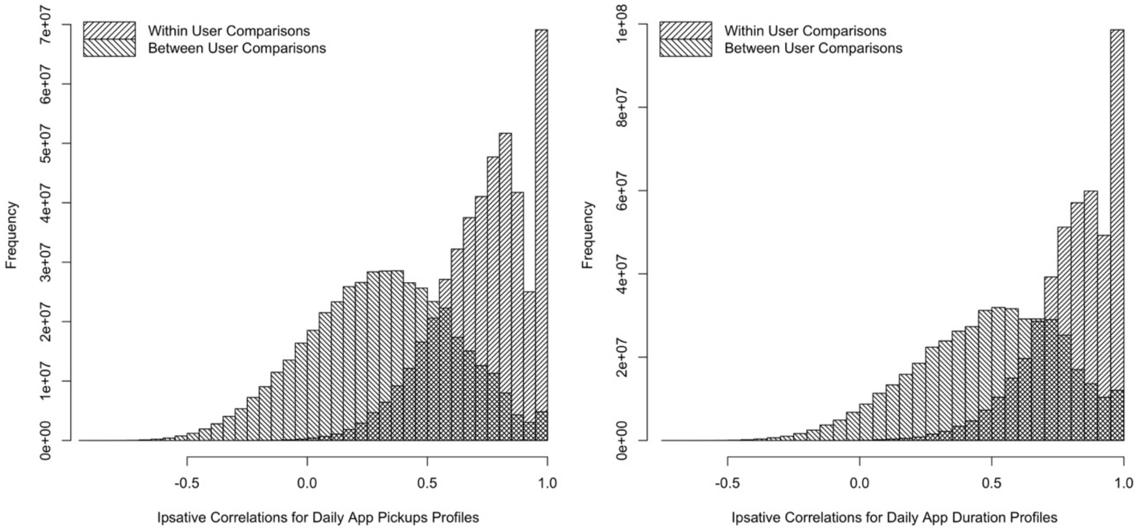

Figure 1. Ipsative correlation coefficients as a function of within-user comparisons and between-user comparisons for Pickups (left panel) and Duration (right panel). A higher coefficient represents greater similarity in the profile of behavior across the comparison pair. The graph style replicates Shoda, Mischel, and Wright (1993) as a tribute to their work.

To assess the robustness of our analysis, we ran two complementary tests. First, because both within-user and between-user distributions deviated from normality, we ran a nonparametric comparison using Wilcoxon rank-sum test and Vargha and Delaney's *A* effect size.[1] These analyses replicated our finding that within-user comparisons were significantly more consistent than between-user comparisons for Pickups, $W = 88{,}324{,}600{,}000{,}000$, $p < 0.001$, $VD.A = 0.12$, and for Duration, $W = 85{,}210{,}000{,}000{,}000$, $p < 0.001$, $VD.A = 0.15$. Second, we re-analyzed the data using a 'split half' comparison, creating an average behavioral profile for the first half and second half of a user's data and comparing these for a within-user comparison, or comparing one half to another user's half for a between-user comparison. This split half approach removes the unbalanced influence that users with more behavioral profiles have in the day pair comparisons, since all users have only two data points. As before, Pickups were significantly higher in within-user comparisons ($M = 0.89$, $SD = 0.10$, *95%CI* [0.88, 0.90]) compared to between-user comparisons ($M = 0.23$, $SD = 0.30$, 95%*CI* [0.21, 0.26]), $t(942.70) = 57.13$, $p < 0.001$, $d = 2.89$, *95%CI* [2.75, 3.03], and Durations were significantly higher in within-user comparisons ($M = 0.91$, $SD = 0.09$, *95%CI* [0.91, 0.92]) compared to between- user comparisons ($M = 0.39$, $SD = 0.32$, *95%CI* [0.37, 0.41], $t(905.43) = 43.46$, $p < 0.001$, $d = 2.20$, *95%CI* [2.07, 2.33].

Identifying Individuals from App Use

Given the intraindividual stability in daily app use, one practical question is, to what extent can a user be identified within a crowd of data based on historic information. This has important security and privacy applications, such as identifying people across multiple devices (e.g., burner phones). Classification algorithms were used to explore this question of profile 'uniqueness'. To do this, each user became a class in a categorical variable, which had

[1] Vargha and Delaney's *A* show the proportion of times Condition A will be higher than Condition B. It ranges from 0.00 – 1.00 with 0.50 representing equal probability (i.e., no difference) and values moving toward 0.00 and 1.00 indicating less equivalence.

780 classes (users). Therefore, the aim of this analysis was to build models which could predict which user was associated with each daily profile.

Random forest models were our classification algorithm of choice. This was because building models with a high number of classes is computationally intensive, and algorithms such as neural networks could not be trained on our high-end cluster. However, random forest models are alternatively very efficient, and previous literature showed they have competitive accuracy in comparison to many other classification models (Fernándes-Delgado, Cernadas, & Barro, 2014). Consequently, we trained a random forest model for Pickups and Duration (separately) using the R package 'rpart' (Therneau, Atkinson & Ripley, 2019). The data entered into the models were the behavioral profiles, which contained the 21 normalized application usage scores, per day, per user. As each behavior profile in the data was paired with a user, and a day (e.g., person 10, day 2) we used this information to both train and test the models. Specifically, as all 780 users had at least 7 days of data, we used the first six days of users' profiles to train the models and their seventh day profile as test data. Therefore, training data consisted of 126 data points per person (21 apps and 6 days), and test data consisted of 21 data points per person (21 apps and 1 day).

Both random forests contained 3120 trees (4 x n), each taking a bootstrapped sample of the data and only selecting 4 variables to be assessed per split (mtry = √ 21) when building individual trees. No pruning took place and trees and were grown to full size. When assessing confusion matrices, the Pickup random forest model classified users from their seventh behavioral profile with 35.76% accuracy, *95%CI* [32.4%, 29.25%], NIR[2] = .0013, *p* <.001; the Duration random forest model classified users with 38.46% accuracy, *95%CI* [35.03%,

[2] No Information Rate (NIR)

41.98%], NIR = .0013, *p* <.001. See supplementary data for performance measures for each class (user) including Sensitivity (*M* = .36), Specificity (*M* = 1), and Recall (*M* = .36)[3].

Probabilities that a behavior profile belongs to each user can be exported from the random forest models. Each user can then be ranked for each behavior profile, from the least to the most probable user. As a result, it is possible to assess the classification accuracy of both random forest models when investigating if the correct user appears in the top 10 most probable users. This assessment showed the accuracy rates of our random forest models on test data increased to 73.46% for Pickups and 75.25% for Duration when success was counted as the user appearing in the highest 10 (approximately the top 1%) of probabilities. Therefore, our models show the potential to narrow down a subject pool of 10 individuals from their daily app use data, with a 3 in 4 success rate.

Discussion

It is almost five decades since Mischel (1973) outlined an 'interactionist' conception of behavioral dispositions, yet most evidence for the theory comes from observations of 'offline' interactions. Here we considered consistency in digital behaviors, through studying the variation of engagement (a behavior) across several nominal situations (apps), collected unobtrusively every second across several days. We found that smartphone users have unique patterns of behaviors for 21 different apps and the cues they present to the user. These usage profiles showed a degree of intraindividual consistency over repeated daily observations that was far greater than equivalent inter-individual comparisons (e.g., a person consistently uses Facebook the most, and Calculator the least every day). This was true for the daily duration of app use but also the simpler measure of daily app pickups—how many times you open each app per day. It was also true for profiles derived from individual days and profiles

[3] Performance means across all classes were equivalent across pickups and duration forests.

aggregated across multiple days. Therefore, by adopting an interactionalist approach in personality research, it is possible to predict a person's future behavior from digital traces whilst mapping the unique characteristics of a particular individual. As research indicates people spend on average 4 hours a day on their smartphone and pick them up on average 85 times a day (Ellis, Davidson, Shaw, & Geyer, 2019), it is important that theories can adapt to the way people behave presently, in digital environments.

When examining *if...then* statements, it may be considered a limitation that we did not examine within app behaviors (e.g., posts and comments), that result from experiencing the 'active ingredients' of a particular digital situation. Future research may wish to explore data that can be retrieved from different applications which share similar behaviors (e.g., posts across different social media sites). Instead, we examined the cross-situational engagement (a behavior) with each app (situation), which is comparatively a simple digital trace, which can be collected easily and unobtrusively, to demonstrate that this alone has within-person consistency.

Consequently, the extent to which our daily smartphone use could act as a digital fingerprint, sufficient to betray our privacy in anonymized data or across devices (e.g., personal vs. work phone), is an increasing ethical concern. Our study adds value to the existing literature by illustrating how engagement with applications alone has within-person consistency that can identify an individual. We modelled users' unique behaviors by training random forests, and then used their exported predictions to assign them to a top-10 candidate pool in separate data with 75.25% accuracy. Thus, an app granted access to smartphone standard activity logging could render a reasonable prediction about a user's identity even when they were logged-out of their account. Similarly, if an app received usage data from several third-party apps, our findings show that this can be used to profile a user and provide a signature that is separate from the device ID or username. So, for example, a law enforcement investigation seeking to identify a criminal's new phone from knowledge of their historic phone use could reduce a candidate pool of ~1,000 phones to 10 phones, with a 25% risk of missing them.

Pertinently, this identification is possible with no monitoring of the conversations or behaviors within the apps themselves, and without 'triangulation' of other data, such as geo-location. Perhaps this should come as no surprise. It is consistent with other research that shows how simple meta-data can be used to make inferences about a particular user, such as personality from smartphone operating system used (Shaw et al., 2016), a particular user from installed apps (Tu et al., 2018) and a person's home location from sparse call logs (Mayer, Mutchler & Mitchell, 2016). Given that many websites and apps collect this meta-data from their users, it is important to acknowledge that usage alone can be sufficient to identify a user if misused. It underscores the need for researchers collecting digital trace data to ensure that usage profiles cannot be reverse engineered to determine participants' identities, particularly if data are to be shared widely. Thus, context-dependent intraindividual stability in behavior extends into our digital lives and its uniqueness affords both opportunities and risks.

References

Aledavood, T., López, E., Roberts, S. G. B., Reed-Tsochas, F., Moro, E., Dunbar, R. I. M., & Saramäki, J. (2015). Daily rhythms in mobile telephone communication. *PLoS ONE*, *10*(9), 1–14. doi:10.1371/journal.pone.0138098

Alvarez-Lozano, J., Osmani, V., Mayora, O., Frost, M., Bardram, J., Faurholt-Jepsen, M., Kessing, L. V. (2014). Tell me your apps and I will tell you your mood: Correlation of apps usage with biploar disorder state. *Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments, 19,* 1-7 doi:10.1145/2674396.2674408

Belk, R.W. (2013). Extended Self in a Digital World. *Journal of Consumer Research, 4*(3), 477-500. doi: 10.1086/671052

Davidson, B. I., & Joinson, A. N. (2021). Shape shifting across social media. *Social Media + Society*, *7*(1), 1-11. doi:10.1177/2056305121990632

Ellis, D. A., Davidson, B. I., Shaw, H., & Geyer, K. (2019). Do smartphone usage scales predict behavior? *International Journal of Human Computer Studies*, *130*, 86–92. doi:10.1016/j.ijhcs.2019.05.004

Fernándes-Delgado, M., Cernadas, E., Barro, S. (2014). Do we need hundreds of classifiers to solve real world problems? Journal of Machine Learning Research, 15, 3133-3181. http://jmlr.org/papers/v15/delgado14a.html

Furr, R. M., & Funder, D. C. (2004). Situational similarity and behavioral consistency: Subjective, objective, variable-centered and person-centered approaches. *Journal of Research in Personality, 38,* 421-447. doi: 10.1016/j.jrp.2003.10.001

Harari, G. M., Müller, S. R., Stachl, C., Wang, R., Wang, W., Bühner, M., … Gosling, S. D. (2019). Sensing sociability: Individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life. *Journal of Personality and Social*

Psychology [Online ahead of print]. doi:10.1037/pspp0000245

Mayer, J. Mutchler, P., & Mitchell, J. C. (2016). Evaluating the privacy properties of telephone metadata. *PNAS, 113,* 5536-5541. doi:10.1073/pnas.1508081113

Nouwens, M., Griggio, C.F., Mackay, W.E. (2017). "WhatsApp is for family; Messenger is for friends": Communication Places in App Ecosystems. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems,* 727-735. doi: 10.1145/3025453.3025484

Mischel, W. (1973). Toward a cognitive social learning reconceptualisation of personality. *Psychological Review, 80,* 252-288. doi:10.1037/h0035002

Mischel, W. (2004). Toward an integrative science of the person. *Annual Review of Psychology, 55,* 1–22. doi:10.1146/annurev.psych.55.042902.130709

Mischel, W., & Shoda, Y. (2010). The situated person. In B. Mesquita, L. Feldman Barrett, & E. R. Smith (Eds.), *The mind in context* (pp. 149-173). New York: Guilford.

Shaw, H., Ellis, D. A., Kendrick, L-R., Ziegler, F. V., & Wiseman, R. (2016). Predicting smartphone operating system from personality and individual differences. *Cyberpsychology, Behavior, and Social Networking, 19,* 727-732. doi: 10.1089/cyber.2016.0324

Shaw, H., Ellis, D. A., & Ziegler, F. V. (2018). The Technology Integration Model (TIM). Predicting the continued use of technology. *Computers in Human Behavior*, *83,* 204-214. doi: 10.1016/j.chb.2018.02.001

Shoda, Y., Mischel, W., & Wright, J. C. (1993). Links between personality judgements and contextualized behavior patterns: situation-behavior profiles of personality prototypes. *Social Cognition, 11,* 399-429. doi:10.1521/soco.1993.11.4.399

Shoda, Y., Mischel, W., & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology, 67,* 674–687. doi:10.1037/0022-3514.67.4.674

Therneau, T., Atkinson, B., & Ripley, B. (2019). ‘rpart’. (Version 4.1-15). [R Package]. http://cran.r-project.org/package=rpart

Tu, Z., Li, R., Li, Y., Wang, G., Wu, D., Hui, P., Su, L., & Jin, D. (2018). Your apps give you away: Distinguishing mobile users by their app usage. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2,*138. doi:10.1145/3264948

Wilcockson, T. D. W., Ellis, D. A., & Shaw, H. (2018). Determining typical smartphone usage: What data do we need? *Cyberpsychology, Behavior, and Social Networking, 21,* 395-398. doi:10.1089/cyber.2017.0652