HUMsim: A Privacy-Oriented Human Mobility Simulator

Giurlanda Francesco[1] Perazzo Pericle[1] and Dini Gianluca[1]

University of Pisa, Italy,
[name.surname]@iet.unipi.it

Abstract. Location-based services rise high privacy concerns because they make it possible to collect and infer sensitive information from a person's positions and mobility traces. Many solutions have been proposed to safeguard the users' privacy, at least to a certain extent. However, they generally lack of a convincing experimental validation with real human mobility traces. Large databases of real mobility traces are extremely expensive to build or buy. In this paper, we present *HUMsim* (*Human Urban Mobility Simulator*), a generator of synthetic but realistic human traces oriented to the experimental validation of privacy solutions. HUMsim generates trajectories that reflect possibly privacy-sensitive habits of people and that, at the same time, account for constraints deriving from a real map. We also validate the soundness of the produced traces by statistically comparing them to real human traces.

Key words: privacy, human mobility, simulation

1 Introduction

The number of GPS-equipped smartphones has recently experienced an exponential growth. 850 millions of devices in 2011 became 1.2 billions in 2012 and are expected to get 4 billion in 2018 [10]. This caused an equally considerable proliferation of *location-based* services and applications. The problem with these services is that they are invasive from a privacy point of view. Not only they make it possible to track users, but also to collect traces, analyze them, and discover sensitive information about users' habits.

The scientific community has developed many solutions that strive to provide a viable trade-off between privacy and performance in location-based services. Among them [2, 9, 12]. The problem with most of these solutions is that they generally lack of a convincing experimental validation on *real* human mobility traces. Large databases of real human mobility traces do exist. Unfortunately they do not come for free. Companies and organizations owning such databases sell them at high prices. On the other hand, running in-home experimental campaigns to take real traces is often impractical.

A possible solution is to use synthetic trajectories. The existing mobility models generate trajectories that are similar to the real ones in terms of, for example, speed, direction changes, presence of obstacles and so on. In these

models, waypoints are typically chosen at random. This implies that the resulting synthetic trajectories display good statistics that are useful for mobile network analysis (cellular networks, MANETs, etc.). However, this also implies that they do not reflect the habits of a person and therefore they are hardly useful for the validation of privacy solutions.

We propose *HUMsim* (*Human Urban Mobility Simulator*), a human mobility simulator aimed at the validation of location privacy solutions. HUMsim generates *semantic trajectories* which are sequences of *semantic waypoints*, i.e., locations labelled with semantic tags [1, 4]. Semantic trajectories are generated according to a *behavioral model* of a person, which describes his daily behavior in terms of visited semantic waypoints (which ones and for how long). For instance, the behavioral model describing a smoker contains semantic waypoints describing stops at a smoke shop. In short, semantic waypoints can reveal information about the person's habits that can put at risk his privacy. Furthermore, HUMsim translates semantic trajectories into *raw trajectories*, which take into account real maps. It follows that the resulting trajectories not only represent "realistic" movements of a person but they also convey privacy-relevant information. As such, differently from existing mobility models which generate trajectories without a semantic value, the semantic trajectories produced by HUMsim allow us to validate location-privacy solutions. We evaluate the soundness of our approach by statistically comparing the trajectories generated by HUMsim to real human traces along two dimensions: the radius of gyration and the displacement between consecutive waypoints. In particular, we compare our results with the results in [11], where the trajectories generated by 100,000 individuals in the European territory are examined. We found some affinity with the results in [11], which corroborates the validity of our synthetic semantic trajectories.

The rest of the paper is organized as follows. Section 2 presents relevant related works. Section 3 describes the HUMsim simulator in detail. Section 4 experimentally evaluates the soundness of HUMsim traces. Finally, the paper is concluded in Section 5.

2 Related Works

The analysis and the modeling of human mobility have always been a challenge for scientists of different disciplines. It allows us to optimize many processes which are related to the daily life of many people. At the same time, understanding the human mobility model allows us to reproduce human behavior in new scenarios. Many models have been proposed to approximate the movements of a person. In this section we survey some of them.

Random Walk (RW) model [6] aims at simulating the unpredictable movement of entities in nature. In RW, each node chooses a random speed inside a predefined range $[V_{min}, V_{max}]$, and a random direction.

In Random Waypoint (RWP) model [3], a mobile node begins by staying in one location for a certain period of time. The node then travels towards a new random destination at a random speed in a predefined range $[V_{min}, V_{max}]$. Upon

arrival, the node pauses for a specified time period before starting the process again. The problem of this approach is the clustering of nodes that occurs at the center of the simulation area. This happens because the mobile nodes tend to pass through it to reach other destinations.

Random Direction (RD) model [16] is designed to overcome the clustering behavior of RWP. The RD mobility model lets the nodes choose a random direction, rather than random position, in which to travel similarly to RW.

Markovian Random Path (MRP) model [6] reduces the sudden changes of speed and direction that afflict the previous models. Improvements to this approach have been introduced with Gauss-Markov (GM) [15] and Markovian Waypoint (MWP) [13]. These models are slight variants of previous random models as they implement Markovian transition probabilities among waypoints or prohibit unrealistic abrupt velocity changes.

All these mobility models are useful to describe the movements of particles or other physical entities which follow random paths, but they badly fit the mobility of human entities which are affected by many variables like personal interests, habits, etc.

Self-similar Least Action Walk model (SLAW) [14] is more accurate and realistic than the previous mobility models. SLAW represents inherent social contexts among walkers manifested as common gathering places and walk patterns therein. SLAW can also express the trip patterns present in the daily mobility of humans. People typically keep a routine of visiting the same places every day, such as an office, but at the same time make irregular trips. SLAW uses Least Action Trip Planning (LATP) to calculate the trip sequence among all the selected waypoints. SLAW effectively expresses mobility patterns arising from people with some common interests or within a single community. Relevant examples are students in the same university campus or people in theme parks where they tend to share common gathering places. But on a larger scale, such as the urban environment, the choice of the waypoints is driven by the habits of a person, or the nearest destination to accomplish a task, or the best path to pass through all the planned waypoints. These are choices taken automatically by an actual driver around the city. Thus, the traces generated by SLAW do not reflect a behavioral model of a person and are not suitable for validating privacy solutions.

In realistic situations, the travel pattern of a node is restricted by the city section that is a urban area with a street network. The mobile nodes have to take into account the traffic limitations and avoid obstacles. City Section Mobility model [7] considers these aspects. The simulation takes place in a realistic city section with different kinds of roads. But, even if the generated trajectories are more similar to those of an actual driver, the destination points are still chosen randomly and they do not have any semantic validity.

All these "artificial" mobility models have to be compared with real human traces in order to prove their validity. Recently, the scientific literature approached the analysis of human mobility from real traces, and supposed that humans follow a *Lévy-flight model* [5, 8, 11]. This model foresees many short

movements around a spot, mixed to few long movements. The length of the movements follows a *power-law* probability distribution. Such a distribution is long-tailed, thus it gives a non-negligible probability of long movements. This model has shown to be consistent with large databases of real human traces, measured by means of banknote spending [5], mobile phone calls [11], or location-sharing check-in's [8]. We used the Lévy-flight model in order to validate the soundness of the paths generated by HUMsim.

3 Human Urban Mobility Simulator

HUMsim (Human Urban Mobility simulator) is a generator of synthetic human traces, aimed at the validation of location privacy solutions. HUMsim has the following characteristics:

- During the day, a person follows a *semantic trajectory* that touches some *semantic waypoints* (home, work, shops, etc.). Some of these waypoints can be privacy-sensitive, i.e. they can reveal private habits or other sensitive information about the person. For example, where the person lives, or whether the person visits a hospital specialized on some particular disease, etc. The definition of what is privacy-sensitive and what is not depends on the particular application, and does not fall in the scope of the simulator.
- The waypoints can be *usual* (e.g. home, work, etc.), which are chosen once and never changed for a person, or *opportunistic* (e.g. markets, shops, etc.), which are chosen time-by-time depending on the current position of the person.
- The daily path and the pause times on the various waypoints can change day-by-day in a probabilistic manner, following a *behavioral model*.
- Once the semantic trajectories have been generated, they are translated into *raw trajectories*, which take into account the mobility constraints, the streets, and the travel times.

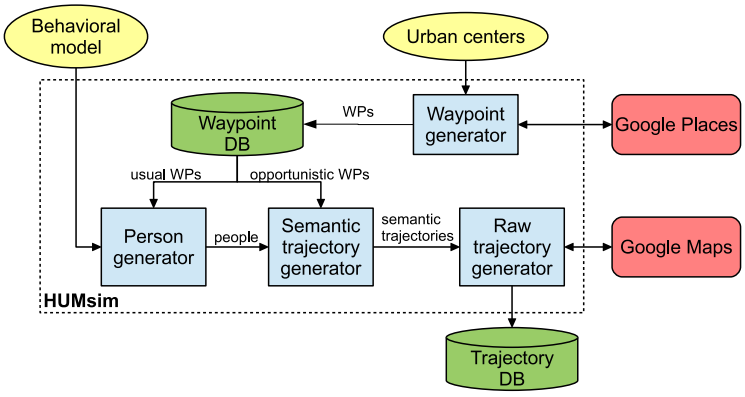


Fig. 1. HUMsim block diagram

HUMsim is composed by four main components, as shown in Figure 1. The *waypoint generator* generates a database of usual and opportunistic waypoints. Depending on their semantic, the waypoints are generated by means of two techniques: *Google Places query* or *random generation*. Some waypoints (typically shops, restaurants, etc.) are generated by means of Google Places query. In this case, the waypoint generator queries the Google Places API by means of a keyword (e.g. "restaurant"). All the other waypoints (typically houses, workplaces, etc.) are generated randomly. The randomly generated waypoints follow a Gaussian distribution centered on a urban center specified by the user. The user can specify more than one urban center, each of which is identified by a circle. The radius of each circle is proportional to the size of the urban center. The generation process first chooses a particular urban center with a probability proportional to its radius, then generates the waypoint according to the Gaussian distribution. Randomly generated waypoints may end up in unaccessible areas such as rural fields or lakes. We use the Google Maps service to move them to the nearest valid locations. The resulting waypoints are stored in a waypoint database.

The *person generator* generates a set of people. A person is represented by (i) a set of usual waypoints, and (ii) a behavioral model. In practice, the person generator randomly selects a set of usual waypoints from the waypoint database, and associates a behavioral model to them. The "home" waypoint is selected for first. The other waypoints are chosen in such a way their distance from "home" follows a power-law distribution. Therefore, a person prefers usual waypoints which are close to home but he does not exclude the possibility of longer distances. We notice that this way of choosing the usual waypoints makes the final raw trajectories more realistic (i.e. closer to a Lévy-flight model, see Section 4).

The *semantic trajectory generator* generates a given number of daily semantic trajectories for each person. A daily semantic trajectory is a sequence of semantic waypoints (W_i) that the person visits in a day, together with a pause time (Π_i) for each semantic waypoint.

$$\text{semantic trajectory} ::= (W_1, \Pi_1) \dots (W_n, \Pi_n)$$

The *raw trajectory generator* translates the semantic trajectory into a raw trajectory, that is an ordered sequence of tuples on the form:

$$\text{raw trajectory} ::= (lat_1, lng_1, t_1) \dots (lat_m, lng_m, t_m)$$

similar to a GPS trace, where lat_i and lng_i are respectively latitude and longitude and t_i is the timestamp associated to the position. While the semantic trajectory considers the movement between two waypoints to be instantaneous, the raw trajectory takes into account also the travel time and path among them.

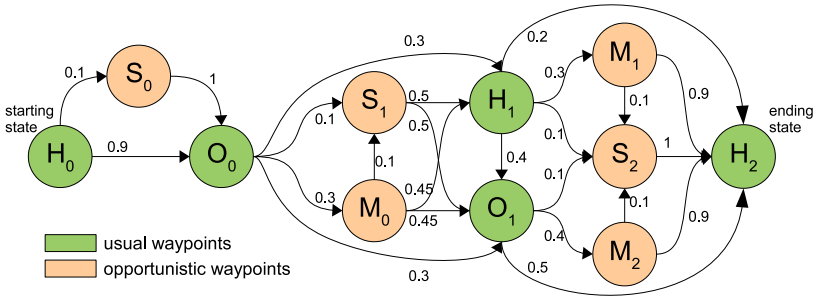


Fig. 2. Example of behavioral model (transition scheme)

State	T_c	T_v	State	T_c	T_v	State	T_c	T_v
H_0	0 min	30 min	O_0	270 min	30 min	M_0	5 min	25 min
H_1	120 min	60 min	O_1	180 min	60 min	M_1	15 min	45 min
H_2	-	-	$S_{0,1,2}$	5 min	5 min	M_2	10 min	30 min

Table 1. Example of behavioral model (pause scheme)

3.1 Behavioral model

The behavioral model specifies the daily mobility of a person in probabilistic terms. It is described by a *transition scheme* and *pause scheme*. The transition scheme is a discrete-time Markov chain, whose states represent distinct visits of semantic waypoints. Figure 2 shows an example of transition scheme.

In this scheme, we modeled a person which moves between four semantic waypoints, "home" (H), "office" (O), "markets" (M) and "smoke shops" (S). Each semantic waypoint can be visited several times during a day. For example, O_0 and O_1 represent same waypoint O that is visited twice in a day.

The pause scheme (Table 1) is a table used to compute the pause time Π_i for each state, that is how long the person waits in that state. This parameter is fundamental for algorithms that analyze the trajectories in order to profile a person, because it gives a quantitative indication of how important the place is for the person. The pause time is composed by two values: a constant value T_c added to a variable value T_v. The T_c is the minimum pause time for that position. The T_v indicates the variation of the pause time.

$$\Pi_i = [T_{c_i}, T_{c_i} + T_{v_i}] \tag{1}$$

The table keeps the values of T_c and T_v for each state. Table 1 shows an example of pause scheme. The example behavioral model of Figure 2 and Table 1 represents a smoker that works and returns home at the end of the day. During the day, the person makes some stops at markets or smoke shops. For each simulated day, the simulator starts in state H_0 and computes a sequence of states that ends in state H_2. In the morning, the person goes to work, possibly visiting a smoke shop (states H_0, S_0, O_0). In the afternoon, he possibly visits a market, or a smoke shop, or both (S_1, M_0), and then he returns to work or goes home

(H_1, O_1). In the late afternoon, he possibly visits again a market, or a smoke shop, or both (S_2, M_1, M_2), and he finally returns to home (H_2). An example of semantic trajectory is:

$$(H_0, 10), (O_0, 275), (S_1, 6), (O_1, 192), (M_2, 15), (H_2, -)$$

where the pause times are expressed in minutes. The semantic waypoint H_2 does not have a pause time because it is the end of the semantic trajectory. This simple behavioral model serves only as a proof of concept, and does not aim to be fully realistic and representative of all people's habits. HUMsim allows the user to define more complex and realistic behavioral models. The user can also specify several behavioral models, describing the daily behavior of different people. When the person generator creates a person with his usual waypoints, it assigns a behavioral model to him for the entire simulation.

3.2 Semantic trajectory and raw trajectory generators

The semantic trajectory generator receives a behavioral model as input and generates a semantic trajectory for each simulated day and for each person. This is done by realizing the Markov chain stochastic process for each day. Then, it assigns a position to each semantic waypoint. As we said before, the positions of the usual waypoints of a person are fixed, decided *a priori* by the person generator. On the contrary, the positions of the opportunistic waypoints are chosen on-the-fly by the semantic trajectory generator, depending on the current position of the person and an *opportunistic choice rule*. HUMsim supports several opportunistic choice rules. For example the "nearest" rule, e.g. choosing the market nearest to the current position, or the "nearest-with-score" rule, e.g. choosing the nearest market having a certain score, etc.

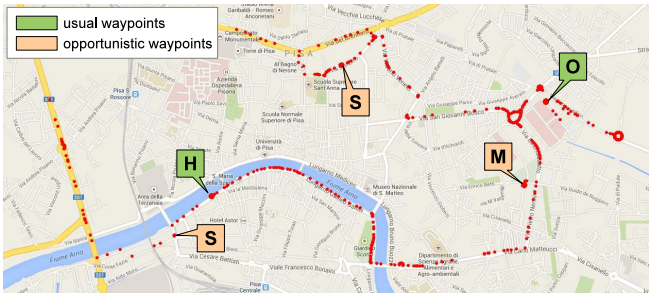
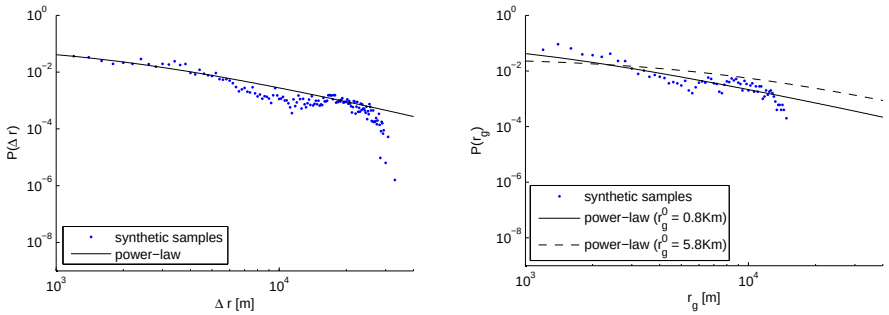


Fig. 3. Example of daily trajectory

The raw trajectory generator translates the semantic trajectory into a raw trajectory. This component uses the positions of the semantic waypoints to calculate the raw trajectory. It leverages on the Direction Service (DS) provided by Google Maps API. The query to the DS needs the list of the visited positions and the mode of transport used. By now, HUMsim implements a single



(a) Displacements (b) Radius of gyration

Fig. 4. Statistical comparison between HUMsim and real human traces

mode of transport for all simulated people (*car*). Future improvements of the simulator will include transport preferences as part of the behavioral model of the person. The DS responds with one or more possible paths and a travel time for each of them. The raw trajectory generator chooses a path and extracts the sequence of points (latitude and longitude) which identify the raw trajectory between two semantic waypoints. Moreover, for each point the raw trajectory generator computes a timestamp abiding by the trip time of the path. During a pause in a waypoint, the trajectory generator continues to produce position samples around the position of the semantic waypoint.

Figure 3 shows an example of daily trajectory of a person which lives in Pisa, Italy. The positions of the semantic waypoints are displayed. Note that the person opportunistically chooses two positions for the smoke shop (semantic waypoint S), depending on his current position.

4 Experimental Validation

We run HUMsim using the example behavioral model of Figure 2 and Table 1, which represents a smoker that works and returns home at the end of the day. Our simulation is focused on the area around Pisa, Italy. We have simulated 5000 users which generate daily trajectories for a period of 30 days. We used the "nearest" opportunistic choice rule for choosing the opportunistic waypoints. We evaluated the soundness of the synthetic traces generated by HUMsim, by statistically comparing them with real human traces. We focused our analysis on two parameters which are at the basis of many studies on human mobility patterns [5, 8, 11]: the *displacement* (Δr) and the *radius of gyration* (r_g). By Δr we indicate the movement of a person from a waypoint to another. In [11], the authors found that the distribution of displacements, recorded over a six-month period for 100,000 individuals in the European territory, follows a truncated power-law with the following shape:

$$\mathrm{P}\left(\Delta r\right) \propto \left(\Delta r - \Delta r_0\right)^{-\beta} e^{-\Delta r/\kappa} \tag{2}$$

where $\Delta r_0 = 1.5km$, $\beta = 1.75$, and $\kappa = 400km$. Figure 4a shows the probability distribution of the displacements generated by HUMsim compared to the distribution found by [11]. It can be seen that the synthetic traces well fit the truncated power-law distribution.

We studied also the distribution of the radius of gyration according to the scale of our simulation. The radius of gyration is an estimation of the general mobility of a person. It is computed in the following way:

$$r_g = \sqrt{\frac{1}{N}\sum_i \|X_i - X_{cm}\|^2} \tag{3}$$

where X_{cm} indicates the *center of mass* of the person's movements. Even in this case, the authors in [11] found that the radius of gyration can be approximated with a truncated power-law:

$$\mathrm{P}\left(r_g\right) \propto \left(r_g - r_g^0\right)^{-\beta_r} e^{-r_g/\kappa_r} \tag{4}$$

where $r_g^0 = 5.8km$, $\beta_r = 1.65$, and $\kappa_r = 350km$.

Figure 4b shows the probability distribution of the radius of gyration over all people generated by HUMsim, compared to the distribution theoretically supposed by [11]. We noticed a discrepancy due to the difference in the scale of the scenario. In fact, in the case of [11] the people move in the whole European continent, whereas our simulations are limited to few urban centers. We found that the theoretic distribution better fits the synthetic data by reducing the parameter r_g^0 to $0.8km$, as shown in Figure 4b. Such a parameter correction lowers the probability of large radii of gyration (cfr. Figure 4b), and thus makes the power-law more suitable to the scale of our simulation scenario.

From these statistical comparisons, we can claim that our synthetic traces well approximate real traces.

5 Conclusion and Future Works

In this paper we presented a human mobility simulator called HUMsim that generates daily trajectories reflecting the habits of a person. The simulator allows us to define a user's behavioral model in terms of semantically annotated waypoints and then generates trajectories passing through those waypoints and accounting for map constraints. The resulting trajectories are realistic as the statistical evaluation showed. Furthermore, given their semantic value, trajectories can also be used to validate the privacy countermeasures aimed at protecting privacy in location-based services.

We leave for future works the comparative analysis with real human traces, for example from public datasets like CROWDAD or the Nokia Mobile Data Challenge.

Acknowledgment

This work has been partially supported by the Tuscany region in the framework of the POR CRO FSE 2007-2013 Asse IV Capitale Umano - project SOcial Sensing (SOS); by the Italian Research Project TENACE (pr. no. 20103P34XC); and by Project PITAGORA, cofinanced by the Regional Government of Tuscany (POR CReO Bando Unico R&S 2012) and the European Regional Development Fund (ERDF).

References

1. Alvares, L.O., Bogorny, V., Kuijpers, B., de Macelo, J., Moelans, B., Palma, A.T.: Towards semantic trajectory knowledge discovery. Data Mining and Knowledge Discovery (2007)
2. Beresford, A.R., Stajano, F.: Mix zones: User privacy in location-aware services. In: PerCom Workshops. pp. 127–131 (2004)
3. Bettstetter, C., Hartenstein, H., Pérez-Costa, X.: Stochastic properties of the random waypoint mobility model. Wireless Networks 10(5), 555–567 (2004)
4. Bogorny, V., Kuijpers, B., Alvares, L.O.: ST-DMQL: A semantic trajectory data mining query language. International Journal of Geographical Information Science 23(10), 1245–1276 (2009)
5. Brockmann, D., Hufnagel, L., Geisel, T.: The scaling laws of human travel. Nature 439(7075), 462–465 (2006)
6. Camp, T., Boleng, J., Davies, V.: A survey of mobility models for ad hoc network research. Wireless communications and mobile computing 2(5), 483–502 (2002)
7. Camp, T., Boleng, J., Davies, V.: A survey of mobility models for ad hoc network research. Wireless communications and mobile computing 2(5), 483–502 (2002)
8. Cheng, Z., Caverlee, J., Lee, K., Sui, D.Z.: Exploring millions of footprints in location sharing services. ICWSM 2011, 81–88 (2011)
9. Dini, G., Perazzo, P.: Uniform obfuscation for location privacy. In: Data and Applications Security and Privacy XXVI, pp. 90–105. Springer (2012)
10. Ericsson mobility report: On the pulse of the networked society. Tech. rep., Ericsson (Jun 2013)
11. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. Nature 453(7196), 779–782 (2008)
12. Gruteser, M., Grunwald, D.: Anonymous usage of location-based services through spatial and temporal cloaking. In: Proceedings of the 1st international conference on Mobile systems, applications and services. pp. 31–42. ACM (2003)
13. Hyytia, E., Lassila, P., Virtamo, J.: A markovian waypoint mobility model with application to hotspot modeling. In: Communications, 2006. ICC'06. IEEE International Conference on. vol. 3, pp. 979–986. IEEE (2006)
14. Lee, K., Hong, S., Kim, S.J., Rhee, I., Chong, S.: SLAW: A new mobility model for human walks. In: INFOCOM 2009, IEEE. pp. 855–863. IEEE (2009)
15. May, P., Ehrlich, H.C., Steinke, T.: ZIB structure prediction pipeline: Composing a complex biological workflow through web services. In: Euro-Par 2006 Parallel Processing, pp. 1148–1158. Springer (2006)
16. Royer, E.M., Melliar-Smith, P.M., Moser, L.E.: An analysis of the optimum node density for ad hoc mobile networks. In: Communications, 2001. ICC 2001. IEEE International Conference on. vol. 3, pp. 857–861. IEEE (2001)